TESTING THE UTILITY OF THE CONSORTIUM FOR THE BARCODING OF LIFE'S TWO 'AGREED UPON' PLANT DNA BARCODES, *matK* AND *rbcL*

By

Ian Michael Cohen

A Thesis
Submitted to the Faculty of the
University of Tennessee at Chattanooga
In Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Environmental Sciences

The University of Tennessee at Chattanooga
Chattanooga, Tennessee

May 2011

TESTING THE UTILITY OF THE CONSORTIUM FOR THE BARCODING OF LIFE'S TWO 'AGREED UPON' PLANT DNA BARCODES, *matK* AND *rbcL*

By

Ian Michael Cohen

Approved:

Joey Shaw Assistant Professor (Director of Thesis)	Ethan Carver Assistant Professor (Committee Member)
Stylianos Chatzimanolis Assistant Professor (Committee Member)	Herbert Burhenn Dean of the College of Arts and Sciences

A. Jerald Ainsworth
Dean, The Graduate School

ABSTRACT

DNA barcoding is the use of short standardized regions of DNA to identify unknown specimens to species. Currently, the zoological community has agreed that *cytochrome oxidase I subunit 1* (*COI*), a mitochondrial gene region, will serve as the barcode region for all animal taxa. Due to oftentimes complicated evolutionary histories of plants, the plant barcoding community has had a much harder time agreeing on a gene region or regions that should be used to barcode the various land plant lineages. This is in large part due to poor reproductive barriers, which allows for chloroplast sharing between closely related species.

In the summer of 2009, the Consortium for the Barcoding of Life's Plant Working Group (CBOL, PWG) announced that portions of two coding chloroplast gene regions (cpDNA), *matK* and *rbcL*, would serve as the DNA barcode for all land plants. This recommendation was accompanied by CBOL's call for continued testing of these two regions, along with other potential gene regions that may prove to be more effective barcode regions, such as noncoding cpDNA regions like *trnH-psbA*.

Originally, this project was focused on the utility of three noncoding cpDNA regions (*trnH-psbA*, *trnL-trnL-trnF*, and *trnS-trnG-trnG*) at identifying species from the genus *Prunus* L. Upon the announcement by CBOL, additional sequence data was generated for *matK* and *rbcL* using the same *Prunus* taxa to determine how well these two regions would delimit species compared to the three noncoding cpDNA regions. In addition to this, sequence data for *matK* and *rbcL* were generated for 27 angiosperm taxa and compared to 34 previously tested noncoding chloroplast gene regions to determine their relative genetic variability. This broader study enabled me to directly compare the genetic variability of these two coding regions to that of noncoding regions.

My results for the broader study demonstrate that *matK* and *rbcL* contain less genetic variability than noncoding regions. Based on the number of potentially informative characters (PIC), *matK* was the 25th most variable region and *rbcL* was the 34th most variable region out of 36 regions tested. This low level of genetic variability in these two regions may make it difficult to identify closely related species. I recommend further study of the 34 previously tested noncoding cpDNA regions to determine their respective utility as plant DNA barcodes. For the direct species identification tests using *Prunus*, I found that no region alone or in combination

was able to discriminate > 50% of species, and noncoding cpDNA regions typically outperformed the Consortium for the Barcode of Life's combination of *matK*+*rbcL*.

DEDICATION

This work is dedicated to my daughter, Ella, who has shown me what unconditional love and true courage are all about.

ACKNOWLEDGEMENTS

Dr. Joey Shaw served as my committee chair, and Drs. Ethan Carver and Stylianos Chatzimanolis served on my committee. Several individuals assisted in the collection of plant material, including J. Shaw, D. Potter, J. Wen, and C. Weeks. The Department of Biological and Environmental Sciences at The University of Tennessee at Chattanooga supported me as a graduate student. My research was funded by the UC Foundation Faculty Research Grant for J. Shaw, the UTC Graduate School and Provost Award, and a grant from the National Science Foundation. I thank J. Shaw for serving as a guide through my long and winding road of research and for pushing me to reach my potential and, at times, exceeding my own expectations and goals, S. Chatzimanolis for supporting me with a generous research assistantship that covered my tuition and living expenses. I would be remised if I did not thank fellow and former graduate students, especially M. Montgomery, for their invaluable friendship and support. Finally, I thank my wife, Allison, and daughter, Ella, most of all for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

TESTING THE UTILITY OF THE CONSORTIUM FOR THE BARCODING OF LIFE'S TWO 'AGREED UPON' PLANT DNA BARCODE REGIONS

Introduction

More than 250 years ago, Linnaeus began his quest to name Earth's living organisms using a hierarchical system that clusters organisms into smaller and smaller groups based on the presence or absence of certain morphological features. This system is still in use today and research to identify and name life on earth continues in earnest. Identification of many plant species can be quite challenging due to the imperfect nature of morphology, i.e. flower color or leaf shape can grade through a species range making it difficult to determine how many species are contained in a particular genus (see Shaw and Small, 2005 for example). While many vascular plant species (e.g., *Acer rubrum* L.), as well as, certain plant groups (e.g. ferns and gymnosperms) are relatively easily identifiable, others require that a suite of morphological characters are present to properly identify (e.g., Asteraceae Bercht. & J.Presl, Cyperaceae Jussieu, and Poaceae (R. Brown) Barnhart). For the majority of the year, these hard to identify plant species lack discriminating morphological characters, such as flowers or fruits, making them unidentifiable and increasing the chance that they will go unnoticed (or undocumented) during a typical field season. However, recent advances in technology have enabled biologists to more easily obtain DNA sequences, which could be used to identify species regardless of time of year.

First proposed by Hebert et al. (2003), DNA barcoding is the use of short DNA sequences (≤ 750 bp) for rapid species identification. The potential impacts of DNA barcoding are far reaching and could benefit a number of related disciplines including: ecology, conservation biology, and population genetics (Tautz et al., 2003; Blaxter et al., 2005; Valentini et al., 2009). Despite widespread criticism (Will and Rubinoff, 2004; Mitchell, 2008; Seberg and Petersen, 2009; Spooner et al., 2009), DNA barcoding efforts have garnered the attention, support, and funding from various organizations including the Alfred P. Sloan Foundation, which has committed $150 million to the Barcode of Life endeavor (Alfred P. Sloan Foundation Website, http://www.sloan.org/program/7). This funding has aided in the creation of two consortia, the Barcoding of Life Database (BOLD) and the Consortium for Barcoding of Life (CBOL). The purpose of these two consortia is to 1) serve as a repository for DNA barcode sequence data, 2) assist in the decision process to choose gene regions that will serve as DNA barcodes a gene region or regions, (3) improve and standardize the analytical methods by which the gene regions can be compared to one another to ensure that the best region(s) are selected, and (4) development of a probabilistic search algorithm, so that researchers can compare sequences from unknown samples to known samples.

According to Kress and Erickson (2008), in order for a gene region to be considered as a potential DNA barcode it must meet the following three criteria: (1) contain significant species-level genetic variability and divergence, (2) possess conserved flanking sites for universal primer development for wide taxonomic application, and (3) have a relatively short sequence (≤ 750 bp) in order to facilitate current capabilities of DNA sequencing (this would also facilitate amplification and sequencing from older preserved material since DNA degrades over time).

Currently, zoologists have identified a portion of the mitochondrial *cytochrome oxidase subunit 1* gene (*COI*) to serve as their "universal" barcode (Hebert et al., 2004; Ward et al., 2005), with several other regions suggested for taxa that are more difficult to discriminate (e.g., tardigrades and nematodes; see Meyer and Paulay, 2005 for details). In plants, it is well understood that the evolutionary rate of the mitochondrial genome (mtDNA) tends to be highly conserved (Clare, 2008; Fazekas et al., 2008), thus this genome does not offer enough genetic variability to discriminate between closely related species. Plant mitochondrial genomes also tend to experience high rates of structural rearrangements and gene duplication events, making it difficult to design universal primers that will work across the various land plant lineages. Thus far, identifying a plant DNA barcode region that will work as well as *COI* in animals has eluded the plant barcoding community. Researchers have turned their attention to the nuclear and chloroplast genomes, both of which accumulate mutations faster than the mitochondrial genome does in plants.

The nuclear genome might seem to be the obvious choice since it accumulates mutations faster than either the chloroplast or the mitochondrial genome. Early on, the nuclear ribosomal internal transcribed spacer region (nrITS) was suggested for plant barcoding since it is the most widely sequenced region in plants (Kress et al., 2005) and there is already a framework in place to build a DNA barcode database, however, the nuclear genome has several functional limitations. Because it is biparentally inherited, individuals can be heterozygous at a particular locus. This could make it more difficult to capture genetic variation within a species since an individual can possess two different copies of a gene. Nuclear genes are also often present in multiple copies, so using a region from the nuclear genome could drive the costs of plant barcoding up because many more alleles would have to be sequenced in order to correctly

measure (and associate) variation within a single species. For these reasons, the plant barcoding community has focused on the chloroplast genome (cpDNA).

Like the mitochondrial genome, cpDNA is effectively haploid and non-recombinant. This makes cpDNA a better choice than the nuclear genome because there are no issues related to heterozygosity. cpDNA also accumulates mutations at least three times faster than the plant mitochondrial genome (Wolfe et al., 1987; Palmer et al., 2000), so there are more potentially informative characters to identify closely related species. However, since cpDNA accumulates mutations slower than *COI* does in animals researchers have had to test large portions of the chloroplast genome in order to find gene regions that contain high levels of interspecific variation to differentiate between closely related species. It has become widely accepted that no single chloroplast region contains enough genetic divergence to discriminate the ~ 300,000 species of land plants (Kress and Erickson, 2007; Taberlet et al., 2006; Fazekas et al., 2009), and that plant DNA barcoding is going have to be a multilocus approach. Several multilocus barcode (MBC) approaches have been proposed (Chase et al., 2005; Newmaster et al., 2006; Chase et al., 2007; Kress and Erickson, 2007; Fazekas et al., 2008), but none of these have been shown to have the level of success of *COI*. Fazekas et al. (2009) recently demonstrated that species discrimination only improved slightly when more than three loci were concatenated, which suggests that potential barcode regions need to contain high amounts of genetic variability between species since simply adding more sequence data does not mean improved species delimitation. While the plant barcoding community has narrowed its focus to the chloroplast genome, the debate on whether to use coding or noncoding gene regions continues (Kress et al., 2005; Devey et al., 2009; Ford et al., 2009; Fazekas et al., 2010).

There are advantages and disadvantages associated with using either coding or noncoding cpDNA regions. Noncoding regions are less evolutionary constrained (see Kelchner, 2000), thus mutations can accumulate faster with virtually no effect on phenotype. This could mean more informative nucleotide characters to distinguish between closely related species. Noncoding regions also contain insertion-deletion characters (indels), which have been shown to help resolve phylogenetic relationships at lower taxonomic levels (Gielly and Taberlet, 1994). On the other hand, those championing the use of coding regions argue that current sequencing technology, sequence alignment algorithms, and database search algorithms makes the use of noncoding regions too difficult for barcoding purposes (Ford et al., 2009). This is largely due to the presence of polynucleotide runs in noncoding regions, which can make obtaining quality sequences and aligning sequences more difficult than coding regions (Devey et al., 2009; Ford et al., 2009). However, Kress et al. (2005) argue that sequence alignment issues should not be considered in a barcoding scheme since technology will continue to improve and resolve these issues.

For those in the plant barcoding community advocating the use of noncoding regions, previous work by Shaw et al. (2005; 2007) provides a foundation. Shaw et al. (2005; 2007) tested the relative genetic variability of 34 noncoding cpDNA regions across ten phanerogam lineages. Based on Shaw et al. (2005), and their own empirical testing of nine noncoding regions, Kress et al. (2005) proposed using nrITS and the *trnH-psbA* intergenic spacer from the chloroplast genome. As noted above, nrITS currently presents many challenges for barcoding and has therefore been largely avoided by the plant barcoding community. However, *trnH-psbA* was found to be one of the most variable noncoding cpDNA regions in terms of percent variability (Shaw et al., 2005). This is an attractive barcode region because it has a relatively

short length (avg. aligned length 465 bp; to Shaw et al., 2005), which makes amplification and sequencing generally straightforward. *TrnH-psbA* does have functional limitations that complicate its use as a barcode region, such as poly- or mononucleotide repeats. These repeat regions result in *trnH-psbA* being too short (< 400 bp) in many plant lineages, thus reducing the number of variable nucleotide characters present, or too long in other plant lineages (> 1000 bp) making it more costly to sequence (Chase et al., 2007). Nucleotide repeats also reduce the clarity and reliability of DNA sequences, which can make it exceedingly difficult to successfully sequence the entire region regardless of size (Devey et al., 2009). For these reasons, many in the barcoding community have advocated the use of coding cpDNA regions.

Ford et al. (2009) note that coding regions are easily alignable and the amino acid code can be utilized to ensure that a nuclear pseudogene was not unintentionally sequenced, which they argue outweigh the use of noncoding regions. Using the tobacco chloroplast genome (*Nicotiana tabacum* L.), Ford et al. (2009) looked at 41 of the 81 coding regions to determine which regions would be most suitable for DNA barcoding. After further testing 12 loci on 98 samples, representing the major land plant lineages (mosses, liverworts, gymnosperms, and angiosperms), they determined that six of the twelve regions merited further evaluation based on amplification success and sequence variability, including *matK*, which has been touted early on in the search for a plant DNA barcode (Ford et al., 2009). However, none of the regions tested by Ford et al. (2009) were shown to be as variable as the noncoding regions tested by Shaw et al. (2005; 2007).

Since the publication of Kress et al. (2005), numerous articles have discussed the advantages and disadvantages of a number of other cpDNA regions. But none of these studies have unequivocally shown any chloroplast gene regions to be as effective for plant barcoding as

COI has for animals. Table 1 summarizes the plant DNA barcode studies since 2005 and shows the taxonomic groups used and gene regions tested. These previous studies can be broken down into two types, (1) those that have looked at a suite of coding and noncoding regions using a few species from a variety of genera and families (Fazekas et al., 2008; Ford et al., 2009; Hollingsworth et al., 2009) and (2) those more narrowly focused on testing how well different gene regions identify species within a single family or genus (Ziegenhagen et al., 2005; Edwards et al., 2008; Newmaster et al., 2008; Nitta, 2008; Seberg and Petersen, 2009; Spooner et al., 2009; Newmaster and Ragupathy, 2009). Sampling in many of the latter studies was also geographically restricted. Both types of studies are important for identifying and measuring the utility of potential plant DNA barcodes. Studies with a broad taxonomic scope are useful for determining which regions will amplify and sequence easily across the land plant lineages, while the studies with a narrow taxonomic scope are useful at determining the fine scale utility of each potential plant DNA barcode region at discriminating closely related species. In either type of study, it is possible that intra- and interspecific variation was not accurately captured for a particular species due to a lack of multiple samples per species (see Fazekas et al., 2008) or because sampling did not mirror the geographic distribution for a particular family, genus or species (see Newmaster et al., 2008). It is paramount that within, and between, species genetic variability is measured for a species prior to setting up an efficient barcoding system since genetic variability is backbone of DNA barcoding. Zhang et al. (2010) argue that most DNA barcoding projects have wildly under-sampled at the species level, thus have not measured the genetic variation thoroughly enough to correctly define and identify a species. They also point out that sampling is idiosyncratic, i.e., the number of samples needed per species is unique to that particular species (Zhang et al., 2010); couple this with that fact that most species in BOLD are

represented by fewer than 10 sequences (Hajibabaei et al., 2007) and the chances are high that many species searches might return a misleading species assignment for an unknown sample.

The PWG (2009) looked at seven candidate loci (three coding and four noncoding cpDNA regions) across 907 samples representing all the major land plant lineages to determine which loci amplified the best, produced the cleanest bidirectional reads, and discriminated the most species. They reduced their dataset to only samples that were successfully sequenced for all seven loci tested, so species identification tests were performed on 397 samples (all angiosperms) rather than the original 907 samples (PWG, 2009). It was formally announced in the *Proceedings of the National Academy of Sciences* (CBOL PWG, 2009) that the barcoding community had "agreed upon" the use of portions of two coding cpDNA regions, *matK* and *rbcL*, as the genetic markers to serve as the barcodes for all land plants. As the PWG noted (2009), *matK* is one of the fastest evolving coding regions (Hilu et al., 2003). However, developing universal primers for *matK* has been difficult and the PWG (2009) noted their own difficulties with generating *matK* sequence data for non-angiosperm taxa. On the other hand, *rbcL* is the best characterized cpDNA gene, but not the most variable region, and its inclusion had more to do with its ease of amplification and sequence recovery rather than its ability to discriminate species. The PWG (2009) did find that species resolution was better when *rbcL* was included in multilocus tests. Both regions have been tested in combination with other regions prior to the PWG's announcement and the results appear mixed (Kress and Erickson, 2007; Fazekas et al., 2008; Hollingsworth et al., 2009), with neither *matK* nor *rbcL*, alone or in combination, having been shown to be as successful as *COI* is in animal taxa.

It is well established that most chloroplast gene regions will identify an unknown sample to genus (see Newmaster et al., 2006; Kress and Erickson, 2007), but the goal of DNA barcoding

is species level identification. Large-scale studies with wide phylogenetic/taxonomic scope, such as the PWG (2009), are important in determining amplification and sequencing success of the various regions tested. These types of studies included only a few congeneric or conspecific samples, which may alleviate many of the issues facing plant DNA barcoding, including the potential overlap in genetic distances between closely related species due to hybridization or chloroplast sharing. According to Fazekas et al. (2009), plant barcoding is going to continue to be a challenge due to incomplete lineage sorting and high rates of chloroplast sharing between closely related species. Unlike animals, many plant groups display poor reproductive barriers, with many species are indeed capable of forming interspecific hybrids and in some cases intergeneric hybrids (Rieseberg, 1997). Since species boundaries tend to be porous in plants, defining what is a species can be challenging (Shaw and Small, 2004; 2005) and can add to the difficulty of using DNA barcodes for species identification or association. This lends to the idea that plants are just harder to barcode than animals (Fazekas et al., 2009).

Study Objective

The over arching objective of my research was to compare the utility of *matK* and *rbcL* as plant DNA barcodes to other noncoding cpDNA regions. Previous research looked at the utility of a number of gene regions in one of two manners 1) broad phylogenetic scope (many families or genera with few species) or 2) narrow phylogenetic scope (one family or genus with many species). There are currently no protocols for study design within the barcoding community, so I tested *matK* and *rbcL* in the above two manners to follow the study designs from previous DNA barcoding studies (Table 1).

Using the same samples and methods of Shaw et al. (2005; 2007), the first part of this research was broad in phylogenetic scope in order to measure amplification, sequencing, and

alignability success for *matK* and *rbcL* across angiosperm lineages. Within this framework, comparisons between closely related species were used to measure the relative genetic variability of *matK* and *rbcL* across seven angiosperm lineages. The relative genetic variability *Matk* and *rbcL d*ata were then compared to 34 previously tested noncoding cpDNA regions by Shaw et al. (2005; 2007) to get an idea of how well these two coding regions may identify closely related species.

The second part of my research was narrow in phylogenetic scope to better assess the ability of five cpDNA gene regions to identify species. Sequence data were generated for the following five gene regions: *matK* and *rbcL* (the 'agreed upon' barcoding loci) along with *trnH-psbA, trnL-trnL-trnF,* and *trnS-trnG-trnG*, three noncoding regions that were promoted in earlier studies as potential barcoding regions (Kress et al., 2005; Taberlet et al., 2006; Nitta 2008). Figure 1 shows the location of these five gene regions within the chloroplast genome. Direct species discrimination was performed using 203 accessions of 54 species of *Prunus* L. (Rosaceae). *Prunus* is a large, complex genus with a global distribution and it is known to contain both closely and distantly related species (see part two for further discussion). Each region was tested alone and in all possible multilocus combinations to determine which region or regions delimited the most *Prunus* species.

Materials and Methods

Taxon Sampling

Comparing genetic variability of* matK *and* rbcL *to 34 noncoding gene regions using seven angiosperm lineages—This dataset contained 27 samples representing 27 species from seven lineages of angiosperms (magnoliids, monocots, caryophyllids, eurosids I and II, and euasterids I and II), based on APG II (2003). These are the same accessions that were previously used by

Shaw et al. (2005; 2007) to determine the relative utility of 34 noncoding chloroplast regions for low-level systematics, except for *Minuartia cumberlandensis* (B.E. Wofford and Krall) McNeill because all extracted DNA had been exhausted for previous studies. This sample was replaced by an accession that was shown in an unpublished study to contain no sequence divergence from the original accession (pers. comm. R. Small, University of Tennessee, Knoxville, TN, USA). A list of the taxa used and GenBank accession numbers are provided in Table 1 in Appendix I.

Prunus *species and outgroup selection for testing barcoding efficacy of* matK *and* rbcL *in comparison to three noncoding cpDNA regions*—*Prunus* L. is an excellent model taxon for DNA barcoding because it is a complex, hyperdiverse genus that includes 1) many widespread species (e.g., *P. serotina* Ehrh. and *P. arborea* (Blume) Kalkman, are found throughout eastern North America or Southeast Asia, respectively), 2) several narrow endemic species (e.g., *P. geniculata* Harper and *P. ramburii* Boiss., with the first restricted to the area around Orlando, FL, USA, and the latter Spain), 3) many closely related species (see Shaw and Small, 2004; 2005), and 4) distantly related species as recent molecular work shows that it also includes the previously recognized genera *Pygeum* Gaertn. and *Maddenia* Hooker and Thompson (Potter et al., 2007; Wen et al., 2008). Using *Prunus* s.l. will enable us to better gauge how well the five potential DNA barcodes will work at discriminating species in a large, taxonomically challenging group.

Prunus also includes the agriculturally important fruits like plums, apricots, peaches, cherries, and almonds. According to Boriss et al., (2009), the value of the above mentioned fruits to the U.S. economy in 2007 ranged from $101.1 million for plums to more than $2 billion for almonds. Based on data from the Food and Agricultural Organization (FAO) of the United Nations website (FAOSTAT TradeSTAT website, faostat.fao.org), the collective global

import/export value for fruits in this genus, exceeded $13 billion in 2006. The genus also contains one commercially important medicinal species, *P. africana* (Hook. f.) Kalkm., which is used to treat benign prostatic hyperplasia (BHP) in men over 50 (Stewart, 2003). The estimated annual trade value of *P. africana* is around $220 million and growing rapidly (Cunningham et al., 1997). Because of its global agricultural and medicinal value, the National Clonal Germplasm Repository (NCGR) at Davis, CA has been tasked with the responsibility of collecting, preserving, evaluating, and distributing the genetic resources to ensure that crop diversity in these species is available for future generations (NCGR at Davis website, http://www.ars.usda.gov/main/site_main.htm?modecode=53-06-20-00). NCGR currently houses over 1000 accessions of various *Prunus* species, which are preserved through vegetative propagation either by root cuttings or budding onto rootstocks (NCGR at Davis website).

Today, *Prunus* is thought to include ~ 150 species (pers. comm. Joey Shaw, The University of Tennessee at Chattanooga, Chattanooga, TN, USA) found mainly in the temperate regions of the Northern Hemisphere, with scattered distributions found throughout the tropical latitudes in southeast Asia, sub-Saharan Africa, and South America (Lee and Wen, 2001). According to Lee and Wen (2001), the generic delimitation of *Prunus* has been controversial starting with de Tournefort's work in which he used fruit morphology as the primary means for sorting the various taxa into six distinct genera. Since the publication of de Tournefort's work, the classification scheme for *Prunus* species has been amended at least ten times with as few as two genera recognized within *Prunus* s.l. and as many as seven (Lee and Wen, 2001). It was not until 1865 that Bentham and Hooker united the six genera of de Tournefort into *Prunus* s.l., which they further subdivided into seven sections (Lee and Wen, 2001).

Rheder's (1940) classification scheme, in which he included roughly 120 species or lesser taxa, is the most widely accepted treatment for the genus. Rheder (1940) subdivided the genus into five subgenera (*Prunus* L., *Amygdalus* L., *Cerasus* Pers., *Padus* (Moench) Koehne, and *Laurocerasus* Koehne.) and 12 sections. Krussman (1978) followed up on this work and included the same five subgenera, but went a step further and split section *Microcerasus* out of subgenus *Cerasus* Pers. creating a sixth subgenus, *Lithoceraus* Ingram. While *Prunus* taxonomy has been controversial, recent molecular work has largely supported the classification schemes of Rheder (1940) and Krussman (1978) (see Bortiri et al., 2001; 2002; and 2006; Shaw and Small 2004; 2005; Wen et al., 2008).

Our dataset contained 203 accessions representing 54 species of *Prunus* collected from around the world. All five subgenera recognized by Rheder (1940), *Prunus* s.s. L., *Amygdalus* L., *Cerasus* Mill., *Padus* Mill, and *Laurocerasus* M. Roem., were included along with four accessions from the previously recognized genus *Pygeum.* Of the 54 species, 48.1% were represented by two samples, 42.6% were represented by three to six samples, and 9.3% were represented by seven or more samples. *Prunus armeniaca* L. and *Prunus serotina* Ehrh. were sampled the heaviest (12 samples each). *Physocarpus opulifolius* (L.) Maxim. served as the outgroup taxon for analyses based on phylogenetic monophyly and was collected from a natural population in Tennessee. It was used as our outgroup taxon because none of the taxa within *Prunus* has unequivocally been shown to be sister to the rest of the genus (Mowrey and Werner, 1990; Bortiri et al., 2001; 2002; Lee and Wen, 2001).

Ingroup sampling of *Prunus* was from wild-collected populations (collected by J. Shaw, D. Potter, J. Wen, or S.-W. Chin) or leaf material sent to us by Clay Weeks from the National Clonal Germplasm Repository at Davis, California, U.S.A, with several exceptions. *Prunus*

maritima Marshall var. *gravesii* (Small) Watson, which has likely been extirpated from the wild, was obtained from the University of Connecticut, Storrs, Connecticut, USA. Samples of *P. subcordata* Benth., were sent to us by colleagues at Oregon State University, Corvallis, Oregon USA. Positive identification of wild-collected samples were made by experts in this genus including: D. Potter, J. Shaw, S-W Chin, or J. Wen. Wild-collected specimens were identified using the following works: Small (1933), Rehder (1940), Bailey and Bailey (1941), Fernald (1950), Blackburn (1952), Gleason (1952), Steyermark (1963), Radford et al. (1968), Correll and Johnston (1970), Duncan and Duncan (1988), Godfrey (1988), Wunderlin (1988), Gleason and Cronquist (1991), and Flora of North America (1993+, Draft Copy), Smith (1994), Wofford and Chester (2002), and Lingdi et al. (2003). A list of species used in this study, along with collector information and GenBank accession numbers, is provided in Appendix II, Table 1**.**

Marker Selection

Comparing genetic variability of* matK *and* rbcL *to 34 noncoding gene regions using seven angiosperm lineages—Sequence data for *matK* and *rbcL* were generated for 27 angiosperm species to determine how variable these two regions are. Results were subsequently compared to 34 noncoding regions, previously tested by Shaw et al (2005; 2007), to determine which regions contained the highest levels of genetic variability and the greatest potential to resolve relationships between closely related taxa. Laboratory and data analysis protocols are provided below.

Prunus *species and outgroup selection for testing the efficacy of* matK *and* rbcL *in comparison to three noncoding cpDNA regions*—The 203 samples used in this investigation were successfully amplified for the following regions: *trnG*^{UUC} and *trnL*^{UAA} introns, *trnS*^{GCU}-

trnG^{UUC}, *trnL*^{UAA}-*trnF*^{GAA}, *trnH*^{GUG}-*psbA* intergenic spacers, and the portions of the *matK* and *rbcL* genes that were identified by PWG. *trnH*^{GUG}-*psbA*, *trnL*^{UAA}-*trnL*^{UAA}-*trnF*^{GAA}, and *trnS*^{GCU}-*trnG*^{UUC}-*trnG*^{UUC}, were selected for study prior to the PWG announcement (2009), and are based on past DNA barcoding studies (Kress et al., 2005; Taberlet et al., 2006; Nitta, 2008), as well as, the work of Shaw et al. (2005; 2007).

***trnH*^{GUG}*-psbA* (*trnH-psbA*)**—While Shaw et al. (2005) ranked this region in the bottom third of noncoding cpDNA regions due to its short average length (465 bp, Shaw et al., 2005), they did note that it was the second most variable region based on percentage and that it amplified and sequenced easily across all phanerogam lineages. According to the PWG (2009), several members strongly advocated for its inclusion into any multilocus barcoding approach as it had been shown to contain high levels of genetic variability between species and it was found to be very informative in that study and previous ones (Kress et al., 2005; Kress et al., 2009). Several issues have come up regarding the inclusion of *trnH-psbA* as a barcode region including: sequence length variability and mononucleotide repeats. It has been shown that *trnH-psbA* ranges in size from ~ 400 bp to > 1000 bp (Shaw et al., 2005; Kress et al., 2005; PWG, 2009). Current sequencing technology is able to successfully read 750-800 bp of DNA, so any barcode region should be within this range to avoid issues with obtaining quality bidirectional sequence data. Also, *trnH-psbA* contains mononucleotide repeats, which can make PCR amplification and sequencing more difficult since these repeat regions tend to cause slip-strand mispairing (Devey et al., 2009). This disrupts normal PCR amplification by causing the *Taq* polymerase to slip and incorrectly amplify the target sequence (Devey et al., 2009) and can make it costlier to obtain quality DNA sequence data since automated sequencers end up stuttering and misreading the bases downstream of these repeat regions (Devey et al., 2009). While cost is not the only issue or

factor in identifying a plant barcode, it is an important one to consider since any barcoding system needs to be affordable to be successful. We decided to include *trnH-psbA* in this study because it has been touted as a potential plant DNA barcode region early on (Kress et al., 2005) and it has been tested in 13/20 studies (see Table 1, Part 1).

***trnL*^{UAA}*-trnL*^{UAA}*-trnF*^{GAA} (*trnLLF*)**—The *trnL* intron and the *trnL-F* intergenic spacer, alone or in combination, have been two of the more widely used regions for early phylogenetic studies (Shaw et al., 2005), and a search on GenBank (accessed 1/4/2010) yielded over 44,000 accessions. This existing framework may make it tempting to establish a barcoding database centered on this region. However, Shaw et al. (2005) showed that *trnLLF* is a relatively slowly evolving region and a poor choice for low level taxonomic studies. Taberlet et al. (2006) looked at the utility of using the *trnL* intron as a barcode region and found species discrimination was poor. We decided to include this region along with the flanking *trnL-trnF* intergenic spacer, which is easily amplified and sequenced together with the *trnL* intron, to further evaluate the utility of this region as a plant barcode because the intergenic spacer typically tends to be more variable than the *trnL* intron.

***trnS*^{GCU}*-trnG*^{UUC}*-trnG*^{UUC} (*trnSGG*)**—Shaw et al. (2005) showed this region to be among the top tier for low-level phylogenetic studies. Currently, only Nitta (2008) has tested this region in a barcoding context. His results suggest that this is a good potential barcode region since it had the lowest level of intraspecific variability and the highest level of interspecific variability of the three regions he tested (Nita, 2008). Low intraspecific variation and high interspecific variation are important components of a barcode region. Low levels of genetic variability present in a single species means fewer individuals will be needed to capture intraspecific genetic variation,

while a high interspecific variation means that it will be more likely that closely related species can be delimited. However, Nitta (2008) did not outright advocate the use of *trnSGG* due to difficulties associated with DNA amplification and aligned sequence length (> 1400 bp in Hymenophyllaceae), which is well over the proposed 750-800 bp limit for DNA barcodes. Like *trnH-psbA*, this region contains mono- and polynucleotide repeats, which can make obtaining quality sequence data difficult. This has required the development of internal primers to ensure quality bidirectional sequencing, which were used in this study for several problematic taxa. Despite these issues, this region was included because of its discriminatory power at low taxonomic levels (Shaw et al., 2005; Nitta, 2008), and because of the low degree of intraspecific variation shown by Nitta (2008).

PWG* matK and rbcL *Barcode Regions—Recently proposed by the PWG, the utility of *matK* has been of much debate (PWG, 2009). According to the PWG (2009), this region is among one of the most rapidly evolving protein coding regions in the chloroplast genome and it has consistently showed high levels of species discrimination in angiosperms. *MatK* was among the first regions suggested as potential plant DNA barcode region (Chase et al., 2007), and this was one of six loci that Ford et al. (2009) strongly suggested for further study due to its ease of amplification and overall high rate of variability. In the PWG (2009) study, only 66% of taxa were correctly identified using *matK* alone and species identification success only improved to 72% in multilocus tests.

Like *trnLLF* and *matK*, *rbcL* was utilized early on in phylogenetic studies and proved useful at the family level (Newmaster et al., 2006). The PWG (2009) noted that *rbcL* was not a highly variable site (alone it identified 61% of taxa), and Ford et al. (2009) noted that the inclusion of *rbcL* in any barcoding scheme had more to do with its historical use in deep-level

phylogenetics and not empirical DNA barcode testing. Its inclusion by the PWG (2009) had more to do with its performance in multilocus tests (72% of species were delimited when combined with *matK*). *MatK* and *rbcL* were included in this study to directly compare the two proposed plant barcode regions to three noncoding regions at species identification.

Laboratory procedures

DNA was extracted from leaves using the DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA). Samples were amplified by polymerase chain reaction (PCR) using the primer pairs listed in Table 2 for *matK*, *trnSGG*, *trnH-psabA*, *trnLLF*, and *rbcL*. PCR was performed using Eppendorf Mastercycler gradient or Mastercycler personal thermal cyclers in 50 μL volumes with the following reaction components: 2 μL template DNA (10–100 ng), 13 *ExTaq* buffer (PanVera/TaKaRa, Madison, Wisconsin, USA), 200 μmol/L each dNTP, 3.0 μmol/L $MgCl_2$, 0.1 μmol/L each primer, and 1.25 units *ExTaq* (PanVera/TaKaRa). Reactions included bovine serum albumin at a final concentration of 0.2 mg/mL, which improved amplification of difficult templates.

All PCR protocols described below were preceded by template DNA denaturation at 80°C for 5 min except for *matK* and *rbcL*, which had an initial denaturation at 94°C for 1 min and 95°C for 4 min respectively. The PCR cycling conditions for *trnLLF* were: 30 cycles of denaturation at 94°C for 1 min, primer annealing at 50°C for 1 min, primer extension at 72°C for 2 min A final extension step consisted of 5 min at 72°C. The PCR cycling conditions for *trnH-psbA* were: 30 cycles of denaturing at 94°C for 30 s, annealing at 50°C for 30 s, and extension at 72°C for 1 min The PCR cycling conditions for *trnSGG* were by touchdown method: 15 cycles of denaturation at 96°C for 1 min, primer annealing at 76°C (20.4°C/cycle) for 45 s, primer extension at 72°C for 2 min, and then 30 cycles of denaturation at 96°C for 1 min, primer

annealing at 69.5°C for 45 s, primer extension at 72°C for 2 min The PCR cycling conditions for *matK* were: 35 cycles of denaturation at 94°C for 30 s, primer annealing at 52°C for 20 s, primer extension at 72°C for 50 s, and a final extension step for 5 min at 72°C. The PCR cycling conditions for *rbcL* were: 35 cycles of denaturation at 94°C for 30 s, primer annealing at 55°C for 1 min, primer extension at 72°C for 1 min, and a final extension step for 10 min at 72°C. PCR products were checked on 1% agarose gels before being cleaned with ExoSAP-IT (USB, Cleveland, Ohio, USA). DNA sequencing was performed in house using the same primers used in amplification (Table 2) with the ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction Kit, v. 2.0 or 3.1 (Perkin-Elmer/Applied Biosystems, Foster City, California, USA) and sent to the Molecular Biological Research Facility (Knoxville, TN, USA) for sequence reads on an ABI Prism 3100 automated sequencer except for *matK* and *rbcL*, which were sequenced commercially using Macrogen (Seoul, South Korea). Sequencher 4.7 (Gene Codes, 2007) was used to edit and assemble complementary DNA strands and check for agreement between them. In no cases, did the adjoined complementary strands disagree.

Data analysis

Due to the number of gene regions tested and the number of analyses performed an outline and explanation of electronic files created and used can be found in Appendix D. For all data sets, alignment of DNA sequences was done by eye in MacClade v. 4.06 (Maddison, 2001; Sinauer, Sunderland, Massachusetts, USA) or Mesquite v. 2.72 (Maddison and Maddison, 2009). Variable positions in the data matrices were double checked against the original chromatogram files to make sure that all base calls were correctly called. In all cases, alignment of potentially informative positions was unambiguous. Because indels have been shown to provide

approximately one-third of the potentially phylogenetically informative information in a cpDNA data set (Gielly and Taberlet 1994), they were coded using FastGap (Borchsenius, 2009).

In the *trnSGG* and the *trnH-psbA* datasets, 150 and 11 bases were omitted from analysis respectively, due to polynucleotide runs, since these maybe PCR artifact and not reflective of the phylogenetic history of the group. Also, in the *trnH-psbA* dataset, an additional 177 bases were omitted due to a large center portion of this intergenic spacer being unalignable. *Prunus* taxa used in a previous study by Shaw and Small (2004), were missing the 3′ *trnL* exon, but since exons are generally highly conserved regions, and likely would not add any information, resequencing of those taxa was not attempted.

Comparing genetic variability of* matK *and* rbcL *to 34 noncoding gene regions using seven angiosperm lineages—*MatK* and *rbcL* were compared directly to the previous regions tested by Shaw et al. (2005, 2007) to determine their relative genetic variability. *MatK* and *rbcL s*equence data were generated for three additional lineages used by Shaw et al. (2005) (gymnosperm, *Eupatorium*, *Solanum*). *MatK* sequences for two of the three gymnosperm samples did not produce quality sequence data, but since only data from the *Tortoise and the Hare II* were available for these three lineages (Shaw et al., 2005); we omitted them from analyses in order to maintain a parallel dataset with Shaw et al. (2007). The relative performance of each region was measured by counting the number of indels and substitutions between the two ingroup taxa, as well as between the ingroup taxa and a third species, which served as an outgroup taxon (see Shaw et al., 2005, 2007 for more detailed explanation of this methodology). The number of indels and substitutions were then summed to give a raw potentially informative character (PIC) value, which was then normalized for each region/lineage combination by dividing the number of PICs found within that region/lineage combination by the sum total of PICs found within a

given lineage. For example, in the magnoliid lineage the 21 PICs tallied for *rpl*16 were divided by the 1050 PICs found within that lineage across 34 noncoding regions plus *matK* and *rbcL*. By doing this, we reduced the influence of differing evolutionary rates or distances among the different taxa.

***Assessment of species identification using* Prunus**—Following CBOL guidelines, uncorrected pairwise genetic distances (UpD) were calculated using PAUP v. 4.0b (Swofford, 2002) for taxa with multiple accessions. Our data was made up of 203 samples representing 53 species or lesser taxa and three unknown samples, which were placed in the dataset to try and determine their respective identities. Additional matrices were created in Mesquite 2.72 for all possible multilocus combinations for a total of 31 datasets (including single locus). This was done to determine which multilocus combinations discriminated the most species. Each distance matrix was then imported into Microsoft® Excel and following PWG (2009) standards, species were considered successfully identified if the highest intraspecific distance was lower than any other interspecific distance (e.g. *Prunus africana* had an ingroup distance between 0 and 0.0017426 in the *rbcL* dataset, which means that all interspecific values must be > 0.0017426 in order for it to be considered positively identified).

A second set of analyses were performed on the same 31 datasets using Bayesian Analysis (BA) to determine the number of species or lesser taxa in the dataset that could be recovered as monophyletic (see Fazekas et al., 2008). Only species with multiple accessions were analyzed since single accession taxa could not be resolved as monophyletic. Due to the size of our dataset, the use of online servers was required to reduce the overall computation time of generating data. Bayesian analyses were performed using Parallel MrBayes @ BioHPC v. 3.1.2 (Ronquist and Huelsenbeck, 2003) (available at: http://cbsuapps.tc.cornell.edu/mrbayes.aspx) to

generate posterior probability distribution using Markov chain Monte Carlo (MCMC) methods. No a priori assumptions about tree topology were made. Models of DNA substitution were estimated using MultiPhyl Online v 1.0.6 (Keane et al., 2007; available at: http://distributed.cs.nuim.ie/multiphyl.php); the General Time Reversible model (GTR, Tavaré, 1986) was used for the *rbcL*, *matK* and *trnLLF* datasets, while the Hasegawa, Kishino and Yano 1985 model (HKY85, Hasagawa et al., 1985) was selected for the *trnH-psbA* and *trnSGG* datasets. Four of the five gene regions tested had a gamma rate. Only *trnSGG* dataset had an I+G rate. MCMC process was set to run for 8 000 000 generations with four chains. This number of generations was chosen because it was the greatest number of generations that Parallel MrBayes would run before timing out on the largest datasets. We wanted to keep all parameters equal between datasets to prevent biased results towards smaller datasets, i.e. those with fewer total nucleotide characters, since these datasets may have been able to run for more generations, thus producing more resolved phylogenies compared to larger datasets. Burn-in was estimated visually by plotting log-likelihood values in Microsoft Excel to determine the number of generations that had run before likelihood values reached an asymptote. To calculate the posterior probability of each bipartition a 50% majority-rule consensus tree was constructed from the remaining trees using PAUP v. 4.0b (Swofford, 2002). Species were considered successfully identified if they formed a monophyletic clade with no other species present and had a posterior probability > 50%.

Results

Amplification, Sequencing, and Alignability

For the seven angiosperm lineages, *matK* and *rbcL* were easily amplified by PCR. Amplification and sequencing of three gymnosperm taxa (*Cryptomeria japonica* (L.f.) Don.,

Taxodium distichum (Nutt.) Croom, and *Glyptostrobus pensilis* (Staunton) K. Koch) was attempted but quality *matK* sequence data could not be obtained for *Taxodium distichum* and *Glyptostrobus pensilis*; however, these three gymnosperm taxa and the 27 angiosperms samples were successfully amplified and sequenced for *rbcL*. The three gymnosperms were omitted from *rbcL* analyses to keep datasets parallel.

Within the *Prunus* dataset, the 203 samples and *Physocarpus opulifolius* were successfully amplified and sequenced for the five potential barcoding regions (*rbcL, matK, trnSGG, trnLLF, and trnH-psbA*). For *trnSGG,* about 33% of the samples required additional sequencing with internal primers to ensure that quality sequences were obtained. M*atK* and *rbcL* were the easiest regions to align because there were no poly A/T runs or hard to align indel regions. Alignment of the *trnLLF* dataset was also relatively easy and no data was considered unalignable. In the *trnSGG* and the *trnH-psbA* datasets, 150 and 11 bases were omitted from analysis respectively, due to polynucleotide runs. More importantly, in the *trnH-psbA* dataset an additional 177 bases (more than half of the raw sequence data) were omitted due to a large center portion being unalignable.

Comparing matK and rbcL to 34 noncoding gene regions using seven angiosperm lineages

The average aligned length of *matK* across the seven lineages was 847 bp, and the aligned length of this region ranged from 823 bp (monocots) to 862 bp (*Gratiola*). For *rbcL,* the average aligned length was 577 bp, and ranged from 573 bp (caryophyllids) to 589 bp (*Gratiola*).

MatK averaged 2.5 times more PICs than *rbcL* across the lineages (average PICs = 25.3; 10, respectively). In *matK,* the magnoliids had the fewest number of PICs (13) while the caryophyllids ranked the highest (45). The number of PICs in *rbcL* ranged from four (monocots)

to 23 (*Gratiola*). In *matK*, the normalized PIC value (which is the percentage that each region contributes to the total number of PICs within a lineage) ranged from 1.2 (magnoliids) to 3.3 (*Carphephorus*). In *rbcL*, the normalized PIC value ranged from 0.4 (monocots) to 1.5 (magnoliids). The average normalized PIC value across the seven lineages was skewed heavily in favor of *matK* (2.27; 0.88, *matK* and *rbcL*, respectively).

In order to determine which of the two gene regions were more variable, percent variability was calculated by dividing the number of parsimoniously informative characters by the aligned length (Table 3). Percent variability for *matK* ranged from 1.6 (magnoliids) to 5.3 (*Gratiola*), and for *rbcL*, the range was 0.7 (monocots) to 3.9 (*Gratiola*). *MatK* was on average more variable across the seven lineages than *rbcL* (average percent variability = 2.97; 1.72, respectively).

In Appendix B, Table 8, the number of substitutions, indels, and inversions can be found for *matK* and *rbcL* in each lineage (for the results of the 34 noncoding gene region/lineage combinations see Shaw et al., 2007). *MatK*, substitutions for ingroup sampling ranged from three in both the monocots and *Prunus* to 22 in *Gratiola*, while the number of substitutions between ingroup and outgroup samples ranged from eight (*Hibiscus*) to 35 (caryophyllids). In *matK*, indels for ingroup sampling ranged from zero in the magnoliids to five in *Gratiola*, and the number of indels between ingroup and outgroup taxa ranged from zero (magnoliids, monocots, and caryophyllids) to four (*Hibiscus*). With respect to *rbcL*, there were three lineages in which no substitutions were counted between ingroup samples (magnoliids, caryophyllids, and *Prunus*), while *Gratiola* had the highest number, nine. Between ingroup and outgroup samples in *rbcL* the number of substitutions ranged from three (monocots) to 16 (magnoliids). Indel characters in

rbcL were only counted in *Gratiola*, which had one between the ingroup taxa and one between the ingroup and outgroup taxa.

Figure 2 shows the average normalized PIC values across the gene region/lineage combinations, and in no lineage was *matK* or *rbcL* the most variable locus. *RbcL* was two to three times less variable than *matK* across all lineages, except in the magnoliids. Figure 3 shows the 34 previously tested regions along with the addition of *matK* and *rbcL*. The highest normalized PIC value was observed in *rpl32-trnL*, which had a normalized PIC value of 5.80 and averaged 63.4 PICs across the seven lineages tested. The *trnS-rps4* intergenic spacer was the least variable, with a normalized PIC value of 0.70 and an average PIC value of 7.7 across the seven lineages tested. A closer look at how much potential information *matK* and *rbcL* might yield to DNA barcoding (or low-level phylogenetic studies) shows that *matK* ranked 25th out of 36 regions tested, based on average normalized PIC value across the seven lineages (2.27); This is less than half as potentially informative as *rpl32-trnL*, and there are five gene regions with at least twice the genetic variability found in *matK*. *RbcL* is the third least variable region tested (34th out of 36; avg. PIC value of 0.88 Fig. 3) and is only potentially better than *psbA-3'trnK* (0.77) and *trnS-rps4* (0.7), which are both relatively short regions and are approximately half the size of *rbcL* (261 bp; 273 bp; 577 bp, respectively). There are 27 gene regions with more than twice the genetic variability found in *rbcL* (Fig. 4). What follows is a closer look at the performance of *matK* and *rbcL* compared to the other regions within each of the seven plant lineages (refer to Appendix 1, Table 2; Figs. 1 and 2 and Shaw et al., 2007).

Magnoliids—The raw values within this lineage ranged from seven in *trnS-rps4* to 86 in *trnQ-5'rps16*, and the normalized PIC value ranged from 0.667 to 8.190 in these same gene regions. *MatK* ranked 30/36 gene regions tested (normalized PIC value = 1.238), and *rbcL* 27/36 gene

regions tested (normalized PIC value = 1.524). This was the only lineage in which *rbcL* outperformed *matK* (16 and 13 raw PICs respectively), but both regions were close to the bottom of regions tested in this lineage and 13 regions were at least twice as variable than either region.

Monocots—Two regions, *trnS-trnG* and *5'rps12-rpl20*, are absent from monocots, so only 34 regions could be compared. The raw PIC value within this lineage ranged from four in *rbcL* to 78 in *rpl32-trnL* and the normalized PIC value ranged from 0.45 to 8.72 in these same gene regions. *MatK* was 25/34 gene regions tested and tied with *psbB-psbH* (normalized PIC value =1.676) *RbcL* ranked 34/34 gene regions tested (normalized PIC value = 0.45). *MatK* was about five times less variable than *rpl32-trnL*, while *rbcL* was nearly 19.5 times less variable.

Minuartia—The raw PIC value within this lineage ranged from two in *trnH-psbA* to 89 in *trnQ-5'rps16*, and the normalized PIC value ranged from 0.13 to 5.82. *MatK* ranked 15/36 gene regions tested (normalized PIC value = 2.94), and *rbcL* 35/36 gene regions tested (normalized PIC value = 0.46). With the exception of *rbcL*, this is one of two lineages in which PIC values were generally higher across all the regions tested than in the other lineages (see Shaw et al., 2007). *MatK* had two to three times the number of raw PICs (45) than it did in five of the six other lineages, but *matK* was still half as variable as the top performing region (*trnQ-5'rps16*) in this lineage, while *rbcL* was only better than *trnH-psbA*.

Prunus—The raw PIC value within this lineage ranged from two in *psbA-3'trnK* and *matK-5'trnK* to 62 in *rpl32-trnL*, and the normalized PIC value ranged from 0.24 to 7.36. *MatK* ranked 22/36 gene regions tested, tied with *trnS-trnM* (normalized PIC value = 2.14), and *rbcL* ranked 32/36 gene regions tested, tied with the *trnL* intron and *3'trnK-matK* (normalized PIC value =

0.71). *MatK* was nearly 3.5 times less variable than *rpl32-trnL*, while *rbcL* was more than 10 times less variable.

Hibiscus—The raw PIC value within this lineage ranged from two in *psbB-psbH* to 81 in *3' trnV-ndhC* and the normalized PIC value ranged from 8.80 to 0.22. *MatK* ranked 17/36 gene regions tested, tied with *trnT-trnL* and the *rps16* intron (normalized PIC value = 2.40), and *rbcL* 33/36 gene regions tested, tied with *3'trnK-matK* (normalized PIC value = 0.76). *MatK* was more than 3.5 times less variable than *3'trnV-ndhC*, while *rbcL* was more than 11.5 times less variable.

Gratiola—The raw PIC value within this lineage ranged from 10 in *trnL-trnF* to 125 in *rpl32-trnL* and the normalized PIC value ranged from 0.53 to 6.62. *MatK* ranked 23/36 gene regions tested (normalized PIC value = 2.23), and *rbcL* 29/36 gene regions tested (normalized PIC value = 1.22). This was the other lineage in which PIC values were generally high across the regions tested when compared to the other lineages. *MatK* was about three times less variable than the top-performing region, *rpl32-trnL*, while *rbcL* was about five times less variable.

Carphephorus—Four of the gene regions previously tested are absent in this lineage (*rpoB-trnC*, *trnS-trnG*, *trnD-trnT*, and *psbM-trnD*) due to rearrangements within the genome. The raw PIC value within this lineage ranged from three in *psbA-3'trnK* to 41 in *atpI-atpH*, and the normalized PIC value from 0.45 to 6.15. *MatK* ranked 18/32 gene regions tested and tied with *ycf6-psbM* (normalized PIC value = 3.3) *RbcL* tied with *ndhJ-trnF*, *psbB-psbH*, *trnL* intron, and *trnL-trnL* for 26/32 (normalized PIC value = 1.05). Of the 32 regions tested in this lineage, *matK* was about half as robust as *atpI-atpH*, while *rbcL* was nearly 6 times less variable.

Assessment of species identification using Prunus

To assess species identification in the model taxon *Prunus* we generated sequence data for *rbcL*, *matK*, *trnSGG*, *trnLLF*, and *trnH-psbA*. Table 4 shows that *trnH-psbA* was the shortest gene region with an aligned length of 520 bp, followed by *rbcL* (574 bp), *matK* (815 bp), *trnLLF* (974 bp), and *trnSGG* (1938 bp). *RbcL* had the fewest total parsimoniously informative characters and indel characters (33) across the five gene regions tested, followed by *trnH-psbA* (94), *matK* (62), *trnLLF* (163), and *trnSGG* (301) (see Table 4). Genetic variability was measured by taking the number of parsimoniously informative characters and dividing by the aligned length. *TrnLLF* was the most variable (9.24%), and *rbcL* was the least variable (5.75%), with *trnSGG*, *trnH-psbA*, and *matK* in the middle of these two regions (7.38%; 7.31%; 6.75%, respectively).

Using 203 samples representing 54 *Prunus* species, species delimitation was tested using each gene region individually and in all possible multilocus combinations with the other four gene regions for a total of 31 separate aligned and coded sequence data matrices (Table 5). Each dataset was analyzed using UpD and BA, which have been two of the more popular metrics used to date. The results are broken down by metric below for easier reading.

Uncorrected p-Distance—Table 5 and Figure 4 show the number of species positively identified for each of the 31 possible gene region combinations. *RbcL* identified the fewest species in single locus tests (7/54 species correctly identified; 12.96%), followed by *matK* (13; 24.07%), *trnH-psbA* (16; 29.63%), and *trnLLF* and *trnSGG* (18; 33.33%). In the single locus tests, only three species were identified by all five gene regions (*P. caroliniana*, *P. maritima*, and *P. subcordata*). A full list of species identified by each single locus is provided in Table 6. The PWG (2009) two

loci combination of *matK+rbcL* identified five more species than *matK* alone and the same number as *trnSGG* alone (18). The three-gene region combination of *matK+trnLLF+trnH-psbA* discriminated the most species (27; 50%); interestingly, all five-gene regions combined identified fewer species (23; 27, respectively) than the above mentioned three-gene region combination.

Monophyly using Bayesian Analysis—Table 5 and Figure 4 show the number of species positively identified for each of the 31 possible single and multilocus combinations. Figure 5 shows the five single locus trees with each monophyletic clade highlighted. No assessment of topological differences between the trees generated was made and only the five single locus trees are shown since there was little improvement on species identification as more data were added. Again, *rbcL* identified the fewest species in single locus tests (7/54 species correctly identified, 12.96%), followed by *trnH-psbA* and *trnLLF* (14; 25.93%), *matK* (16; 29.63%), and *trnSGG* (22; 40.74). In the single locus tests, only one species was identified by all five gene regions (*P. subcordata*). A full list of species identified by each single locus is provided in Table 6. The PWG (2009) two loci combination of *matK+rbcL* identified two more species than *matK* alone but four fewer species than *trnSGG* did alone. The four-gene region combination of *matK+rbcL+trnSGG+trnH-psbA* and the five-gene region combination discriminated the most species (26; 48.15%).

Discussion

It has been difficult to reach a consensus on which gene regions to use for plant barcoding due to their complex and often intertwined evolutionary histories. CBOL accepted the proposal of the PWG (2009) in late 2009, but called for more testing of *matK* and *rbcL* by the

plant barcoding community, as well as other potential DNA barcode regions to determine if these gene regions are indeed the best for plant DNA barcoding. The focus of our research was to test these two touted gene regions by comparing their relative genetic variability to 34 noncoding cpDNA gene regions previously studied by Shaw et al. (2005; 2007), and also to compare the performance of these same two regions with respect to species level identification. What follows is a discussion on the utility of *matK* and *rbcL* as 'universal' plant barcodes based on amplification and sequencing success, alignment issues, genetic variability, and direct species identification.

Amplification, Sequencing and Alignability

Much debate has been focused on identifying gene regions that easily amplify, sequence, and easily align using computerized alignment programs across the various land plant lineages (Kress et al., 2005; PWG, 2009). For these reasons, many in the plant barcoding community have advocated the use of coding over noncoding cpDNA regions to ensure amplification with a single set of primers, quality bidirectional sequencing, and easy alignment using global alignment algorithms without having to adjust sequence data by hand (Newmaster et al., 2006; Ford et al., 2009; PWG, 2009).

Using the PWG (2009) primer sets and PCR protocols for *matK* and *rbcL*, we found no amplification issues across the seven angiosperms lineages or within the 203 accessions within *Prunus* plus *Physocarpus opulifolius* and quality sequence data was obtained for all of the abovementioned accessions. As noted earlier, *matK* sequence data could not be generated for two of the three gymnosperms from the *Tortoise and the Hare* dataset (Shaw et al., 2005). The PWG (2009) noted their own difficulties with obtaining quality sequences for gymnosperms without using taxa specific *matK* primer sets. Why gymnosperms have been difficult to sequence using

'universal' primers is likely due to differing evolutionary histories than angiosperms. It is clear that further evaluation of how to barcode this group is needed before the barcoding community can proclaim that a barcode for land plants has been found.

The use of noncoding regions, such as *trnH-psbA*, have been complicated by the presence of mononucleotide runs (poly A/T runs), which cause slippage of the DNA polymerase during PCR causing a 'stutter' effect (Devey et al. 2009). This makes obtaining quality sequences difficult. We found few difficulties related to PCR amplification and obtaining quality sequences was generally not an issue for the three noncoding regions tested using *Prunus*. Despite several mononucleotide repeat regions in the *trnH-psbA* and *trnSGG* datasets, sequencing was successful because of the short length of *trnH-psbA* (~ 300 bp unaligned) and the use of internal primers to ensure quality sequence data for *trnSGG*. Recently, Fazekas et al., (2010), found that two DNA polymerases, Phusion (Finnzymes, Espoo, Finland) and Herculase II fusion (Agilent, Santa Clara, California, USA) regularly improved quality sequence reads through mononucleotide repeats up to 13 bp. Based on these findings, the use of noncoding regions as barcodes should still be considered since they are generally more variable than coding regions, more robust at infrageneric levels (Gielly and Taberlet, 1994, Kelchner 2000), and obtaining quality sequences is becoming easier through technological advances (Fazekas et al., 2010).

Sequence alignment has garnered the attention of many in the barcoding community because automated alignment programs can align coding regions more accurately and faster than noncoding regions since they generally contain very few indel characters (Ford et al., 2009; PWG, 2009). We attempted to align *trnH-psbA* using Clustal X v. 2.0 (Larkin et al., 2007), but computing processing time was slow (~ four days). More importantly, results could not be relied on since they still needed to be adjusted by hand. The use of coding regions comes at a cost since

there are potentially fewer informative characters to facilitate the identification of closely related species. Noncoding regions generally have a higher rate of nucleotide substitutions and it is has been shown that indel characters can help resolve relationships between closely related taxa (Gielly and Taberlet, 1994; Kress, 2007). Alignment of the three noncoding regions tested using *Prunus* was generally straight forward. The only exception was the *trnH-psbA* dataset, in which 177 bases were omitted (more than half of the total base pairs sequenced), in order to align the data. Despite this omission of data, *trnH-psbA* was still as robust as *matK*, more robust than *rbcL*, and nearly as robust as *matK*+*rbcL* at species identification. According to Erickson et al. (2008), current multiple alignment algorithms are insufficient to handle large amounts of data from a noncoding region and the alternative use of pairwise alignments are too slow for large scale database application. While outside the scope of this paper, improvements to sequence alignment algorithms clearly pose a challenge for the bioinformatics community to solve in order to make analysis of data faster and reduce search times on databases like BOLD or GenBank.

The plant DNA barcoding community has focused on identifying gene regions that are easy to PCR amplify, sequence, and align. Based on these criteria, our data supports the selection of *matK* and *rbcL*. Importantly, we had no PCR amplifying issues and almost no difficulties obtaining quality sequence data for the three noncoding regions used in this study. Our biggest obstacle was the alignment of the three noncoding gene regions, which proved to be much more time consuming than *matK* and *rbcL*. However, Kress et al. (2005) argue that technology will continue to advance and solve this problem, thus noncoding regions should continue to be considered as potential plant DNA barcodes rather than being disregarded by the barcoding community for the sake of simple sequencing and alignment.

Comparing matK and rbcL to 34 noncoding gene regions using seven plant lineages

The focus of *The Tortoise and the Hare* series (Shaw et al., 2005; 2007) was to find the most variable cpDNA regions for low-level phylogenetic research. This work continues today with a number of studies focused on finding coding and noncoding cpDNA region that meet the criteria to be considered a DNA barcode (Haider, 2003; Shaw et al., 2005; 2007; Ford et al., 2008). With respect to phylogenetic studies, chloroplast noncoding regions have been shown to be more useful below the genus level since they accumulate mutations faster than coding regions (Kelchner, 2000).

MatK and *rbcL* were suggested early on in the search for a plant DNA barcoding region (Newmaster et al., 2006; Chase et al., 2007) and have been two of the more widely tested coding regions (Newmaster et al., 2006; Newmaster et al., 2007; Chase et al, 2007; Lahaye et al., 2008; Ford et al., 2009; Hollingsworth et al., 2009; PWG, 2009). Comparing these two regions to 34 previously tested noncoding cpDNA regions, we show that five noncoding gene regions are at least twice as variable as *matK* and 27 noncoding gene regions are more variable than *rbcL* (Fig. 3). Based on these results, more attention should be paid to testing noncoding regions despite the potential obstacles in alignment and search algorithms since they overwhelmingly outperformed *matK* and *rbcL* in this part of our study in every measure (Figs. 2 and 3).

DNA barcoding rests on finding regions with sufficient genetic variability between species, but that variability needs to be low within a species (intraspecific variation) so that, if possible, multiple haplotypes per species do not have to be described. With respect to this criterion, *matK* and *rbcL* are good choices as there was very little genetic distance within a species (intraspecific variation); however, they are poor choices since they also contained very

low levels of genetic variation between species (interspecific variation). Figure 3 shows that there are four regions that are more than twice as variable as *matK* and 28 regions compared to *rbcL*, which suggests that the plant barcoding community should continue to test other regions before settling on these two. In light of our results, we wonder if a plant DNA barcode region(s) should be primarily chosen on the basis of alignability at the expense of species discrimination.

Assessment of species identification using Prunus

For direct species identification tests using *Prunus* L., we were unable to achieve the 70% mark touted by the PWG (2009). Using their combination of *matK*+*rbcL* (18/54 species were positively identified). The three noncoding regions did not achieve the 70% mark alone or in combination, but they did identify more species than *matK* (when analyzed using UpD) or *rbcL* (regardless of metric used) in single locus tests (Table 5). Curiously, *trnLLF* and *trnH-psbA* identified fewer species using BA than UpD, while *matK* and *trnSGG* identified more in locus tests (Tables 5 and 6). *RbcL* was the least informative region at species discrimination in single locus tests, regardless of metric used, and Ford et al. (2009) point out that the inclusion of *rbcL* in any barcoding scenario is due to its historical popularity in phylogenetics not empirical testing. However, the PWG (2009) noted that species resolution improved when *rbcL* was included in multilocus tests, but we did not find this to be the case. In fact, we found that *matK* performed as well or better than *rbcL* in all the various multilocus tests in which the other was excluded and that results were mixed when both regions were present (Table 5).

We found little to no species resolution in many of the subgeneric groups of *Prunus*, irrespective of gene region combinations or metric used, due to zero genetic distance or overlaps in intraspecific and interspecific genetic distances between species. For example, there was no genetic distance between six species of North American plums, and the same six species formed

a polytomy using BA in all 31 consensus trees. Our results are in line with others who found poor species identification due to little genetic variation in cpDNA between species (Edwards et al., 2008; Spooner, 2009; Steel et al., 2010). According to Edwards et al. (2008), at least three gene regions will be needed for species identification for *Aspalathus* L. (Fabaceae), a genus endemic to the fynbos of South Africa. Not surprisingly, our results suggest that it will take between eight and ten gene regions to identify 70% of species in *Prunus*, which has a much broader range than *Aspalathus*.

It is noteworthy that species resolution reached an asymptote as the number of gene regions concatenated increased from two to three (Fig. 6) and species resolution only increased marginally as the number of loci concatenated increased from three to four and in several instances it decreased (Table 5, Fig. 6). According to Fazekas et al, (2009), stringing regions together will do little to improve the overall success of plant barcoding and our results concur. It may be that plants are just harder to discriminate than animals due to lower levels of genetic variability in chloroplast markers (Fazekas et al., 2009). This suggests that species identification in *Prunus*, and very likely in other large genera with closely related species, will continue to be difficult regardless of the gene regions used (see Spooner, 2009; Steel et al., 2010). As Meyer and Paulay (2005) point out, DNA barcoding will work well in thoroughly sampled groups with solid taxonomic foundations, but science has only described 10% of the flora and fauna on Earth (Wilson, 2003), which means there are more poorly understood groups than well understood ones.

Our results, coupled with others (Spooner, 2009; Steel et al., 2010), strongly suggests that developing a universal plant barcoding scheme will continue to be a difficult task. Our results do not support the use of the PWG (2009) *matK*+*rbcL* regions as barcodes for *Prunus* since it only

identified 1/3 of the species in our dataset. Even though the three noncoding regions were at least as variable as *matK* (Table 4), they had their own shortcomings. *TrnLLF* was easy to amplify, sequence, and align, but it did not enhance species resolution in multilocus tests, so its utility as a barcode region is probably limited at best. *TrnSGG* was the best performing region, but it is about 50% longer than *matK+rbcL*, harder to amplify, sequence, and align than either of the coding regions. Despite the drawbacks, *trnSGG* should not be ruled out of any barcoding scheme since it contains a high rate of genetic variability and technology will continue to improve making sequencing and alignment difficulties non-issues. *TrnH-psbA* was the most unusual region, in that it contained a high level of genetic variability despite the omission of nearly 50% of the raw sequence data due to alignability issues. Just like with *trnSGG*, if automated sequence alignment algorithms can be improved, then based our results and those of Kress et al. (2005; 2007) we support its inclusion in any barcoding scheme.

We strongly encourage the plant barcoding community to continue to investigate other gene regions, especially those in the small single copy region, which have received little attention from the plant barcoding community, before settling on *matK+rbcL*. We also urge the community to more robustly test the multigene-tiered approach put forth by Newmaster et al. (2006). The tiered approach starts with a coding region that can provide resolution to family or genus followed by one or more noncoding regions, which can be taxa specific, to resolve samples to species (Newmaster et al., 2006). If the criteria are easy amplification and sequencing across the various land plant lineages (including gymnosperms, bryophytes, liverworts, and pteridophytes), then *rbcL* makes the most sense to be the core gene region since it amplified across all 30 samples in our dataset, sequenced without issue (including gymnosperms), and did not require any manual editing or alignment. If, however, the plant barcoding community wants

to focus on angiosperms, which make up approximately 90% of the world's flora, *matK* would be the better choice since it contained a higher level of genetic variation than *rbcL* (Figs. 1 and 2).

Conclusions

Plant DNA barcoding has been anything but easy. Since Kress et al. (2005), a number of studies have been published that have looked at the utility of a number of coding and noncoding cpDNA regions as potential DNA barcodes across the different land plant lineages. Despite all of this work, results from one study cannot be easily compared to one another due differing study designs. For example, many studies have tested a suite of gene regions across a broad phylogenetic set of taxa. It is well known that genera within and between families of plants are phylogenetically nonequivalent, i.e. genera across families may have very different genetic divergence rates depending on the life histories of the various species included in those genera or families. This suggests that results regarding the utility of gene regions from these studies may be overstated since interspecific variation may not have been fully accounted for within the various genera and families they selected (see Fazekas, 2008; PWG, 2009). On the other hand, studies like Newmaster et al., 2006 may also not have accounted for genetic variability due geographic restrictions and small sample size. This means that while a gene region worked well in a particular study the results cannot be assumed to be a universal truth without more robust testing.

In 2009, in an attempt to standardize plant barcoding, CBOL established the combination of *matK* and *rbcL* as the "universal" plant barcodes based on their performance in previous studies, most notably the PWG (2009). From our study, it is clear that *matK* and *rbcL* lack enough genetic variability to identify closely related species in *Prunus* and therefore be

considered the ‘universal’ plant barcodes. Our results, coupled with previous studies (Kress and Erickson, 2007; Edwards et al., 2008, Spooner, 2009; Newmaster et al. 2008), suggests that it is time to refocus our attention to other regions in the chloroplast genome, such as *ndhf-rpl32* and *rpl32-trnL*, which Shaw et al., (2007) found to be two of the most variable noncoding regions. Many more inquiries similar to the ones performed by Edwards et al. (2008), Farrington et al. (2009), Spooner (2009), and Steel (2010), are needed before we settle on a plant DNA barcode system. Better to go slow than learn later on that it could have been better.

Therefore, we conclude that it would be a mistake to settle on *matK* and *rbcL* in a rush to begin building a plant DNA barcode database unless we are willing to settle for ≤ 70% species resolution. We have shown that a number of chloroplast gene regions display greater levels of genetic variation than *matK* or *rbcL*, as we all as amplify and sequence just as easily. We have also shown that species identification, regardless of metric or gene regions used, will continue to be difficult within a large genus that contains many closely related species, and that there is a need for more testing at the generic level to determine which gene regions will best delimit species. Ultimately, plant barcoding may never be as successful as it is in animals due to the unique and challenging evolutionary histories of plants.

Table 1. A sample of DNA Barcoding Studies. For cpDNA regions, coding regions are listed first followed by noncoding regions in alphabetical order. nDNA = nuclear genome; cpDNA = chloroplast genome.

Author (Year)	Taxa Sampled	Gene Regions Tested
Kress et al. (2005)	99 samples representing 99 species in 80 genera and 53 families	**nDNA:** ITS **cpDNA:** *matK*, *atpB-rbcL*, *psbM-trnD*, *trnC-ycf6*, *trnH-psbA*, *trnL-trnF*, *trnK-rps16*, *ycf-psbM*
Newmaster et al. (2006)	10,000 sequences from GenBank representing the various land plant lineages	**cpDNA**: *rbcL*
Newmaster et al. (2006)	40 samples in 3 genera of Myristicaceae	**nDNA:** *UPA* **cpDNA:** *accD*, *rbcL*, *rpoB*, *rpoC1*, *trnH-psbA*
Taberlet et al. (2006)	123 arctic plants samples 72 food industry plant samples	**cpDNA:** *trnL* intron
Sass et al. (2007)	124 samples representing 21 species in 10 out of 11 cycad genera	**nDNA:** ITS **cpDNA:** *accD*, *matK*, *ndhJ*, *rpoB*, *rpoC1*, *ycf5*, *trnH-psbA*
Kress and Erickson (2007)	48 genera (2 species/genus) across land plant lineages	**cpDNA:** *rbcL*, *trnH-psbA*
Edwards et al. (2008)	133 samples representing 82 species of *Aspalathus*, Fabaceae	**nDNA:** ITS **cpDNA:** *trnH-psbA*, *trnT-trnL*
Lahaye et al. (2008)	101 samples representing 18 angiosperm families from Kruger National Park, South Africa	**cpDNA:** *matK*, *atpF-atpH*, *psbK-psbI*, *trnH-psbA*
Fazekas et al. (2008)	251 samples representing 92 species in 32 genera of land plants	**nDNA:** 23S, rDNA **cpDNA:** *matK*, *rbcL*, *rpoB*, *rpoC1*, *atpF-atpH*, *psbK-psbI*, *trnH-psbA*
Nitta (2008)	37 samples representing all 12 species of Filmy Ferns of Moorea (French Polynesia)	**cpDNA:** *rbcL*, *trnH-psbA*, *trnS-G-G*
Lahaye et al. (2008)	> 1600 Orchidaceae samples from Mesoamerica and southern Africa	**cpDNA:** *accD*, *matK*, *ndhJ*, *rbcL*, *rpoB*, *rpoC1*, *ycf5*, *trnH-psbA*
CBOL PWG (2009)	907 samples of land plants, but only 397, angiosperms were used in species identification tests	**cpDNA:** *matK*, *rbcL*, *rpoB*, *rpoC1*, *trnH-psbA*, *atpF-atpH*, *psbK-psbI*
Farrington et al. (2009)	128 samples representing 19 species of *Caladenia* (Orchidaceae)	**nDNA:** ITS **cpDNA:** *matK*, *ndhF-rpl32*, *psbD-trnT*, *psbJ-petA*, *rpl32-trnL*, *trnL intron*, *trnQ-5'rps16*, *3'trnV-ndhC*
Ford et al. (2009)	98 species of land plants	**cpDNA:** *accD*, *matK*, *ndhA*, *ndhJ*, *ndhK*, *rpl22*, *rpoB*, *rpoC1*, *rpoC2*, *ycf2*, *ycf5*, *ycf9*
Hollingsworth et al. (2009)	44 samples representing 26 species in *Inga*, 42 samples representing 17 species in *Araucaria*, 41 samples representing 26 species *Asterella s.l.* (No comparisons between genera were made)	**cpDNA:** *matK*, *rbcL*, *rpoC1*, *rpoB*, *atpF-atpH*, *psbK-psbI*, *trnH-psbA*
Newmaster and Ragupathy (2009)	56 samples of *Acacia* (Fabaceae)	**cpDNA:** *matK*, *rbcL*, *trnH-psbA*

Table 1 continued

Ragupathy et al. (2009)	40 samples representing 8 species of *Tripogon* (Poaceae)	**cpDNA:** *matK*, *rbcL*, *trnH-psbA*
Spooner (2009)	104 samples representing 63 species of *Petota*, Solanaceae	**nDNA:** ITS **cpDNA:** *matK*, *trnH-psbA*
Seberg and Petersen (2009)	131 samples representing 80 of 81 *Crocus* species	**cpDNA:** *accD*, *matK*, *ndhF*, *rpoC1*, *rps8-rpl36*, *trnH-psbA*
Starr et al. (2009)	93 samples representing 3 subgenera of *Carex* (Cyperaceae)	**cpDNA:** *matK*, *rbcL*, *rpoB*, *rpoC1*, *trnH-psbA*
Asahina et al. (2010)	12 samples representing 5 species of *Dendrobium* (Orchidaceae)	**cpDNA:** *matK*, *rbcL*
Chen et al. (2010)	400 samples representing 326 species in 245 genera and 98 families of land plants	**nDNA:** ITS1, ITS2 **cpDNA:** *matK*, *rbcL*, *rpoC1*, *ycf5*, *trnH-psbA*
Kelly et al. (2010)	23 samples representing 11 species in 6 genera of Podostemaceae from Cameroon and Ghana	**cpDNA:** *matK*, *rpoB*, *rpoC1*, *trnH-psbA*
Steele et al. (2010)	70 samples representing 7 species of *Psiguria* (Cucurbitaceae)	**cpDNA:** *ndhC-trnV*, *ndhF-rpl32*, *psbZ-trnM*, *rpoB-trnC*, *rps16-trnQ*

Table 2. Primers used—Includes primer name, forward and reverse sequences, aligned length of the gene region in *Prunus* L., and author. cpDNA = Chloroplast genome; bp = base pairs.

cpDNA Gene Regions Surveyed	Aligned Length of Gene Region (bp)	Primer Name	Sequence (5′-3′)	Source
matK	815	3F_Kim f	CGTACAGTACTTTTGTGTTTACGAG	CBOL PWG, 2009
		1R_Kim r	ACCCAGTCCATCTGGAAATCTTGGTTC	CBOL PWG, 2009
trnSGG	1938	trnS^{GCU}	AGATAGGGATTCGAACCCTCG	Shaw et al., 2005
		3'trnG^{UUC}	GTAGCGGGAATCGAACCCGCATC	Shaw et al., 2005
trnG-G Internal		5'trnG2G	GCGGGTATAGTTTAGTGGTAAAA	Shaw et al., 2005
trnG-S Internal		5'trnG2S	TTTTACCACTAAACTATACCCGC	Shaw et al., 2005
trnH-psbA	521	trnH^{GUG}	CGCGCATGGTGGATTCACAATCC	Tate and Simpson, 2003
		psbA	GTTATGCATGAACGTAATGCTC	Sang et al., 1997
trnLLF	974	TabC	CGAAATCGGTAGACGCTACG	Taberlet et al., 1991
		TabF	ATTTGAACTGGTGACACGAG	Taberlet et al., 1991
rbcL	574	rbcLa_R	GTAAAATCAAGTCCACCRCG	CBOL PWG, 2009
		rbcLa_F	ATGTCACCACAAACAGAGACTAAAGC	CBOL PWG, 2009

Table 3. Comparison of the PWG (2009) Gene Regions—Comparison of the relative utility of *matK* and *rbcL* to one another across seven angiosperm lineages. Avg. = Average; PIC = Potentially informative characters. Avg. = average; L. = length; PIC = potentially informative characters.

Gene Regions	**Avg. Aligned L. Across Lineages**	**Avg. PICs**	**Avg. Normalized PICs**	**Avg. % Variability Across Lineages**
matK	847	25.3	2.27	2.99
rbcL	577	10	0.88	1.73

Table 4. Comparison of five potential DNA barcoding regions—The relative utility of each potential chloroplast gene region based on parsimoniously informative characters and indels, in 203 samples of *Prunus* L. and *Physocarpus opulifolius* (L.) Maxim. Calculations: % informative characters = parsimoniously informative characters/Aligned Length*100; % Indels = Indels/Aligned length*100; bp = base pairs.

Gene Regions	Aligned Length (bp)	Variable Parsimoniously-Uninformative Characters	Parsimoniously-Informative Characters	% Informative Characters	Indel Characters	% Indels
rbcL	574	11	33	5.75	0	0.000
matK	815	80	55	6.75	7	0.859
trnSGG	1938	80	143	7.38	158	8.153
trnLLF	974	45	90	9.24	73	7.495
trnH-psbA	520	38	38	7.31	56	10.769

Table 5. Assessment of species identification using uncorrected pairwise distance and percent monophyly—54 *Prunus* L. species were used to compare species identification success of five chloroplast gene regions and multilocus combinations of those loci using two common barcoding metrics. This table includes number of species identified and corresponding percentages. Information in bold represents the gene region(s) that identified the most species for each metric. Numbers in bold and italics refer to the PWG (2009) combination of *matK*+*rbcL*. bp = base pairs; ID = Identification; UpD = Uncorrected pairwise distance; BA = Bayesian analysis

Gene Region Combinations Tested	**Number of Nucleotide Characters + INDELS (bp)**	**Number of spp. ID (UpD)**	**Number of Monophyletic spp. (BA)**	**% of spp. ID (UpD)**	**% of spp. ID (BA)**
rbcL	574	7	7	12.96	12.96
matK	822	13	16	24.07	29.63
trnLLF	1047	18	14	33.33	25.93
trnSGG	2096	18	22	33.33	40.74
trnH-psbA	577	16	14	29.63	25.93
matK+trnLLF	1396	21	20	38.89	37.04
matK+trnSGG	2918	20	24	37.04	44.44
matK+trnH-psbA	1399	25	21	46.30	38.89
matK+rbcL	***1396***	***18***	***18***	***33.33***	***33.33***
trnLLF+rbcL	1621	22	16	40.74	29.63
trnLLF+trnSGG	3143	20	22	37.04	40.74
trnLLF+trnH-psbA	1624	25	18	46.30	33.33
trnSGG+trnH-psbA	2673	23	25	42.59	46.30
trnSGG+rbcL	2670	20	23	37.04	42.59
trnH-psbA+rbcL	1151	23	19	42.59	35.19
matK+trnLLF+trnSGG	3965	22	25	40.74	46.30
matK+trnLLF+trnH-psbA	**2446**	**27**	22	**50.00**	40.74
matK+trnSGG+trnH-psbA	3495	23	24	42.59	44.44
matK+trnLLF+rbcL	2443	23	19	42.59	35.19
matK+trnSGG+rbcL	3492	22	25	40.74	46.30
matK+trnH-psbA+rbcL	1973	26	22	48.15	40.74
trnLLF+trnSGG+trnH-psbA	3720	23	23	42.59	42.59
trnLLF+trnSGG+rbcL	3717	21	22	38.89	40.74
trnLLF+trnH-psbA+rbcL	2198	24	20	44.44	37.04
trnSGG+trnH-psbA+rbcL	3247	22	25	40.74	46.30
matK+rbcL+trnLLF+trnSGG	4539	21	23	38.89	42.59
matK+rbcL+trnLLF+trnH-psbA	3020	26	22	48.15	40.74
matK+rbcL+trnSGG+trnH-psbA	**4069**	25	**26**	46.30	**48.15**
trnLLF+trnSGG+trnH-psbA+rbcL	4294	23	24	42.59	44.44
matK+trnLLF+trnSGG+trnH-psbA	4542	23	25	42.59	46.30
All Five Regions	5116	23	**26**	42.59	**48.15**

Table 6. Species Identified—*Prunus* L. species identified by each gene region alone. Species in bold were identified by all five regions independently. For species identified using BA, species are listed in the order they appear on the trees in Figure 4. BA = Bayesian analysis; UpD = Uncorrected pairwise distance.

Gene Regions	Species Identified using BA	Species Identified using UpD
rbcL	*P. caroliniana, P. serotina, P. javanica, P. maritime, P. mume, P. pumila, P. serotina,* ***P. subcordata***	***P. caroliniana***, *P. javanica, P. mahaleb,* ***P. maritima***, *P. mume, P. pumila,* ***P. subcordata***
matK	*P. africana, P. malayana , P. oblongum, P. polystachya, P. caroliniana, P. serotina, P. virginiana, P. phaeosticta, P. zippeliana, P. maritime, P. spinosa,* ***P. subcordata***, *P. mume, P. avium, P. mahaleb, P. petunnikowii*	*P. africana, P. avium,* ***P. caroliniana***, *P. malayana,* ***P. maritima***, *P. oblongum, P. petunnikowii, P. polystachya, P. serotina,* ***P. subcordata***, *P. tenella, P. virginana, P. zippeliana*
trnSGG	*P. africana, P. malayana, P. oblongum, P. polystachya, P. brittoniana, P. ovalis, P. oleifolia, P. caroliniana, P. phaeosticta, P. zippeliana, P. serotina, P. virginiana, P. undulata, P. maritima,* ***P. subcordata***, *P. mume, P. spinosa, P. prostrata, P. pumila, P. tomentosa, P. triloba, P. avium*	*P. africana, P. brittoniana,* ***P. caroliniana***, *P. davidiana, P. domestica, P. integrifolia, P. malayana,* ***P. maritima***, *P. oblongum, P. oleifolia, P. phaeosticta, P. polystachya, P. prostrata, P. pumila, P. serotina, P. spinosa,* ***P. subcordata***, *P. triloba*
trnLLF	*P. africana, P. oleifolia, P. caroliniana, P. phaeosticta, P. zippeliana, P. polystachya, P. javanica, P. maritima, P. spinosa, P. prostrata, P. pumila,* ***P. subcordata***, *P. pensylvanica, P. mahaleb*	*P. africana, P. avium,* ***P. caroliniana***, *P. davidiana, P. domestica, P. mahaleb,* ***P. maritima***, *P. pensylvanica, P. polystachya, P. prostrata, P. pumila, P. reflexa, P. spinosa, P. stipulacea,* ***P. subcordata***, *P. undulata, P. virginiana, P. zippeliana*
trnH-psbA	*P. africana, P. javanica, P. phaeosticta, P. zippeliana,* ***P. subcordata***, *P. serotina, P. prostrata, P. tomentosa, P. mahaleb, P. oblongum, P. polystachya, P. spinosa, P. brittoniana, P. oleifolia*	*P. africana, P. brittoniana,* ***P. caroliniana***, *P. domestica, P. mahaleb,* ***P. maritima***, *P. petunnikowii, P. phaeosticta, P. polystachya, P. prostrata, P. reflexa, P. serotina, P. spinosa,* ***P. subcordata***, *P. undulata, P. virginiana, P. zippeliana*

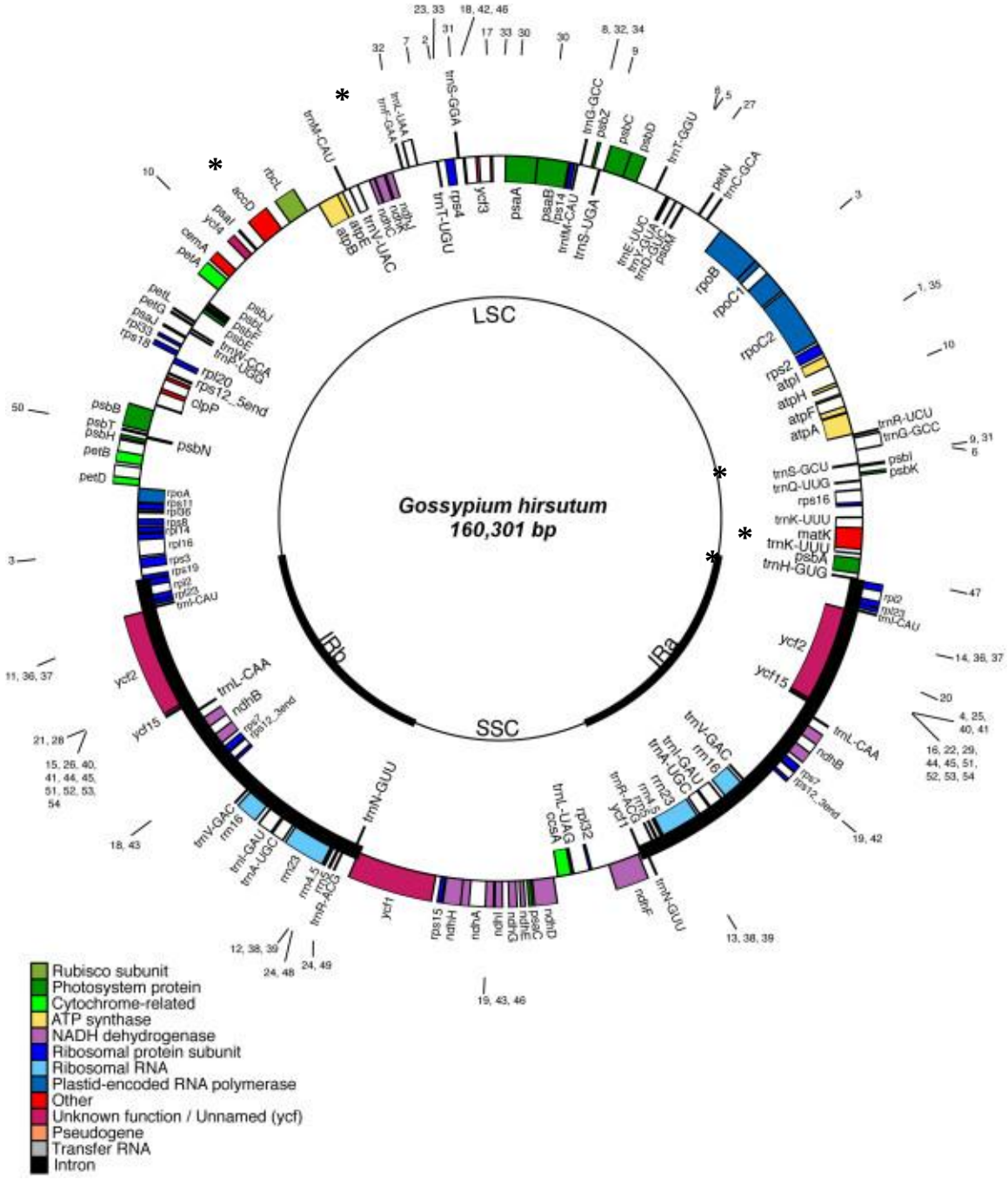

Figure 1. Chloroplast Genome—*Gossypium hirsutum* L. chloroplast genome (Lee et al., 2006) is presented here as a reference map to show where the five gene regions tested are located. Gene regions tested are marked with an *****.

.

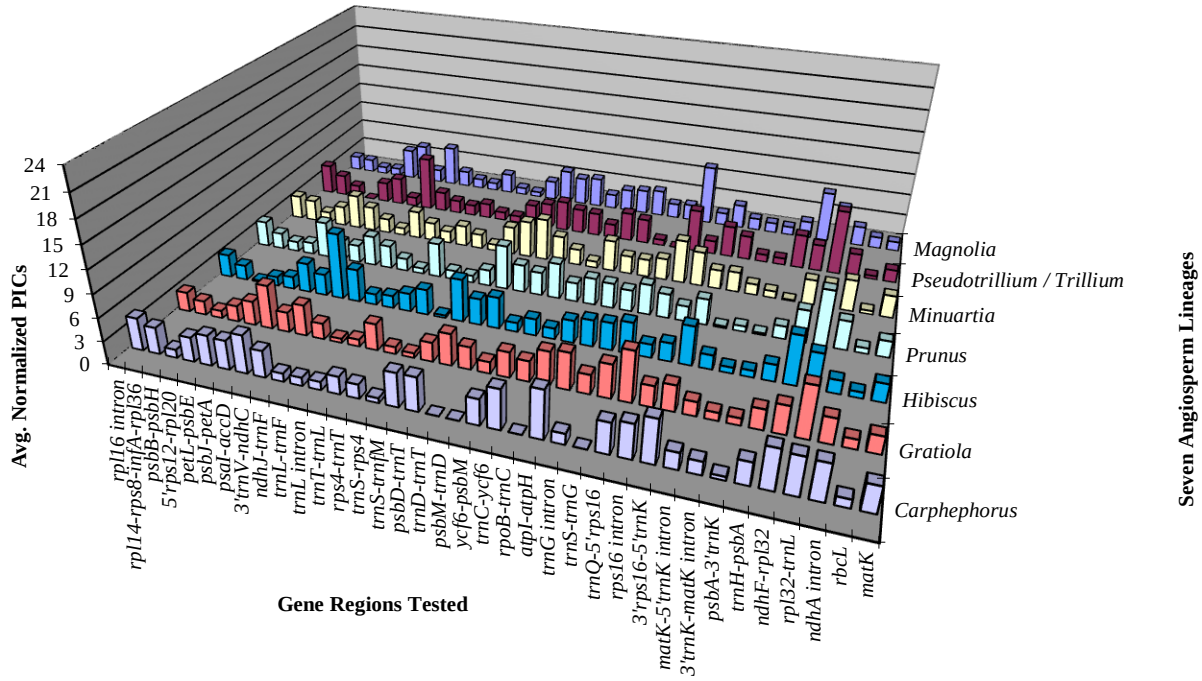

Figure 2. Normalized PIC values across seven angiosperm lineages—Relative genetic variability of 36 chloroplast gene regions across seven angiosperm lineages. Figure recreated using data from Shaw et al. (2007) with *matK* and *rbcL* data added from this study. Avg. = Average; PIC = potentially informative characters

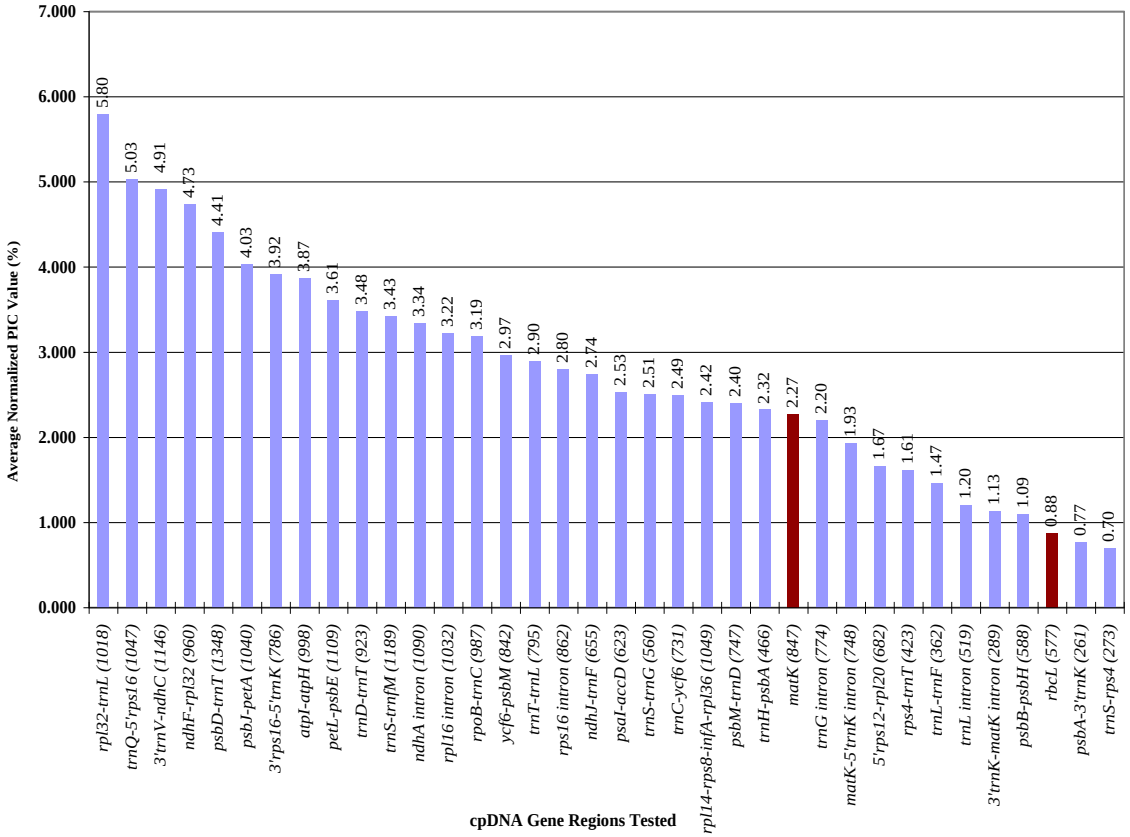

Figure 3. Comparison of *matK* and *rbcL* to 34 noncoding chloroplast regions—Average Normalized PIC values were calculated for the regions below to determine which regions contain the highest levels of genetic variation across seven angiosperm lineages. Figure recreated using data from Shaw et al. (2007). Numbers in parentheses above gene regions represent the average aligned length across the seven angiosperm lineages for that particular gene region. Average aligned lengths may differ from Shaw et al (2005; 2007) due to rounding differences. cpDNA = Chloroplast genome; PIC = Potentially informative characters.

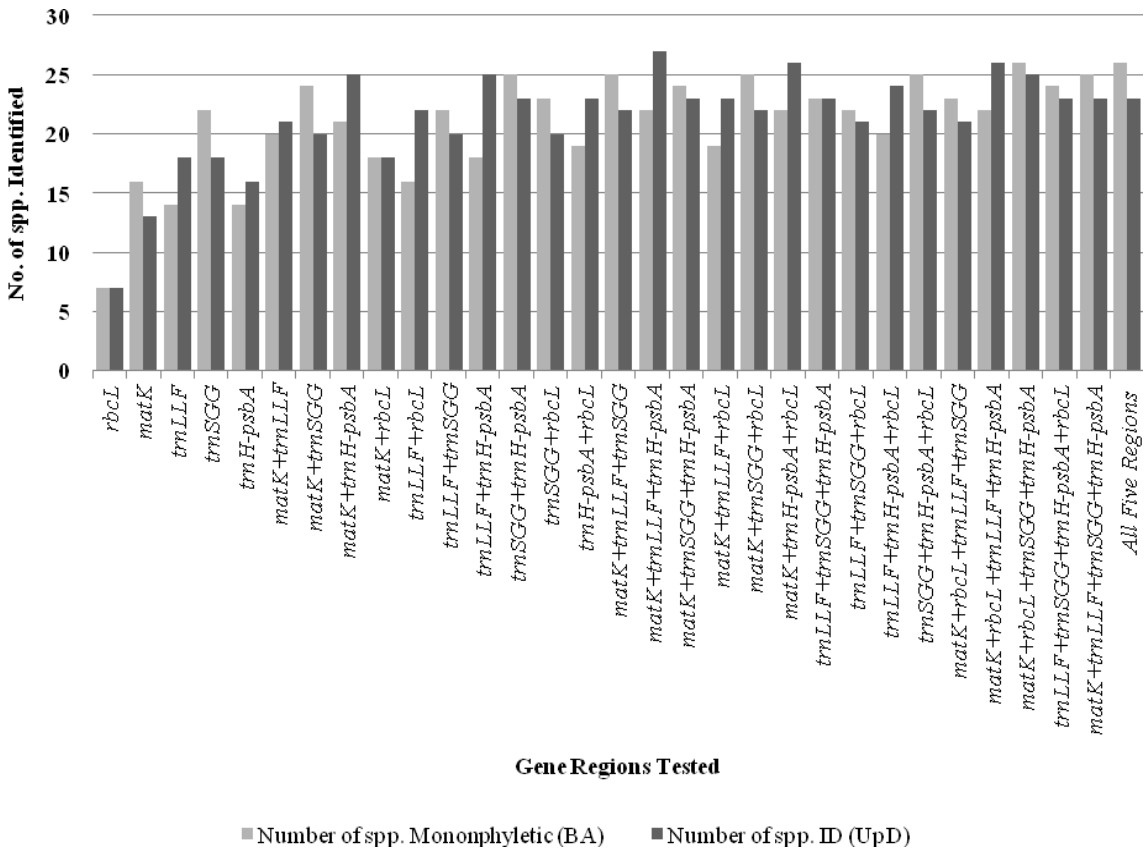

Figure 4. Species Identification of *Prunus*—Percentage of 54 species correctly identified using two common barcoding metrics. spp. = species; ID = identified; UpD = Uncorrected pairwise distance; BA = Bayesian analysis. No. = Number; spp. = species.

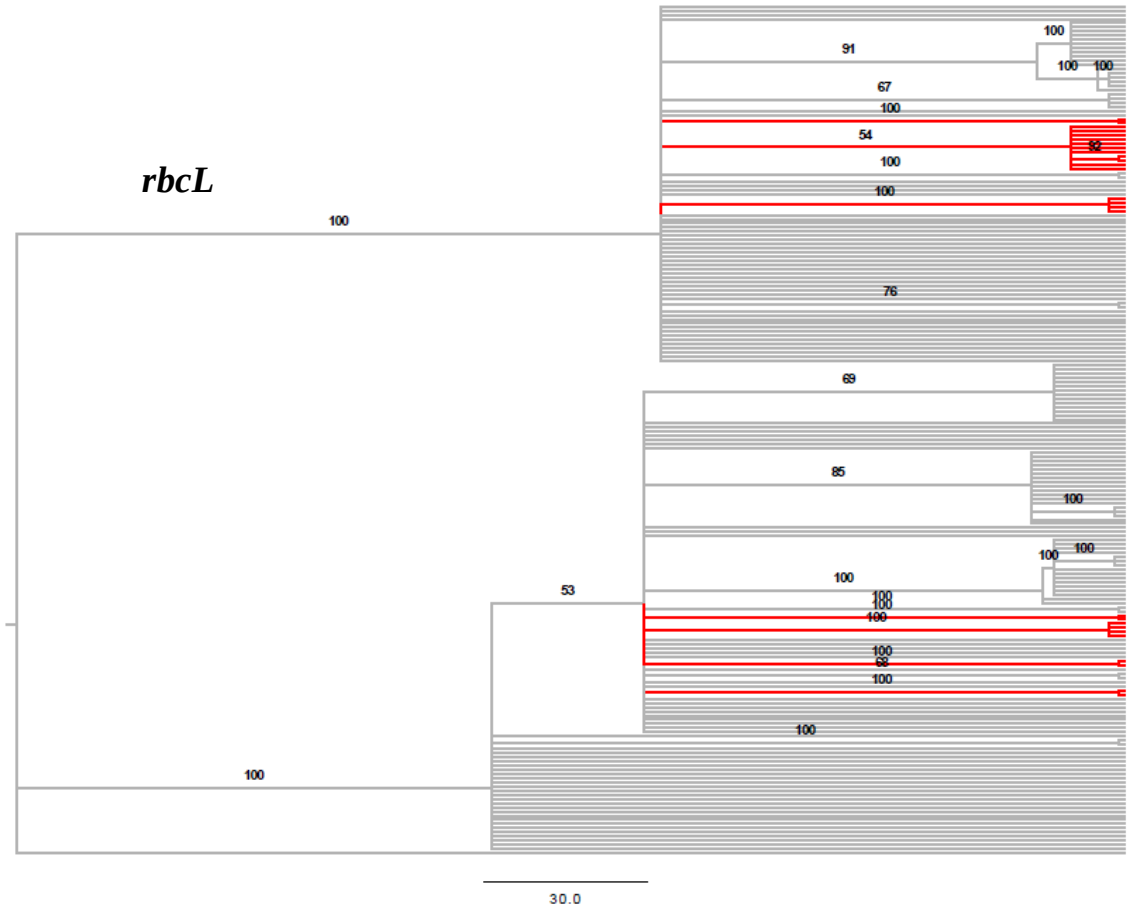

Figure 5. Consensus trees—Species discrimination for the five single loci tests using Bayesian analysis to determine the number of species resolved as monophyletic. Clades highlighted in black were considered correctly identified following Fazekas et al. (2008). Species appear on tress as listed in Table 5 for Bayesian analysis.

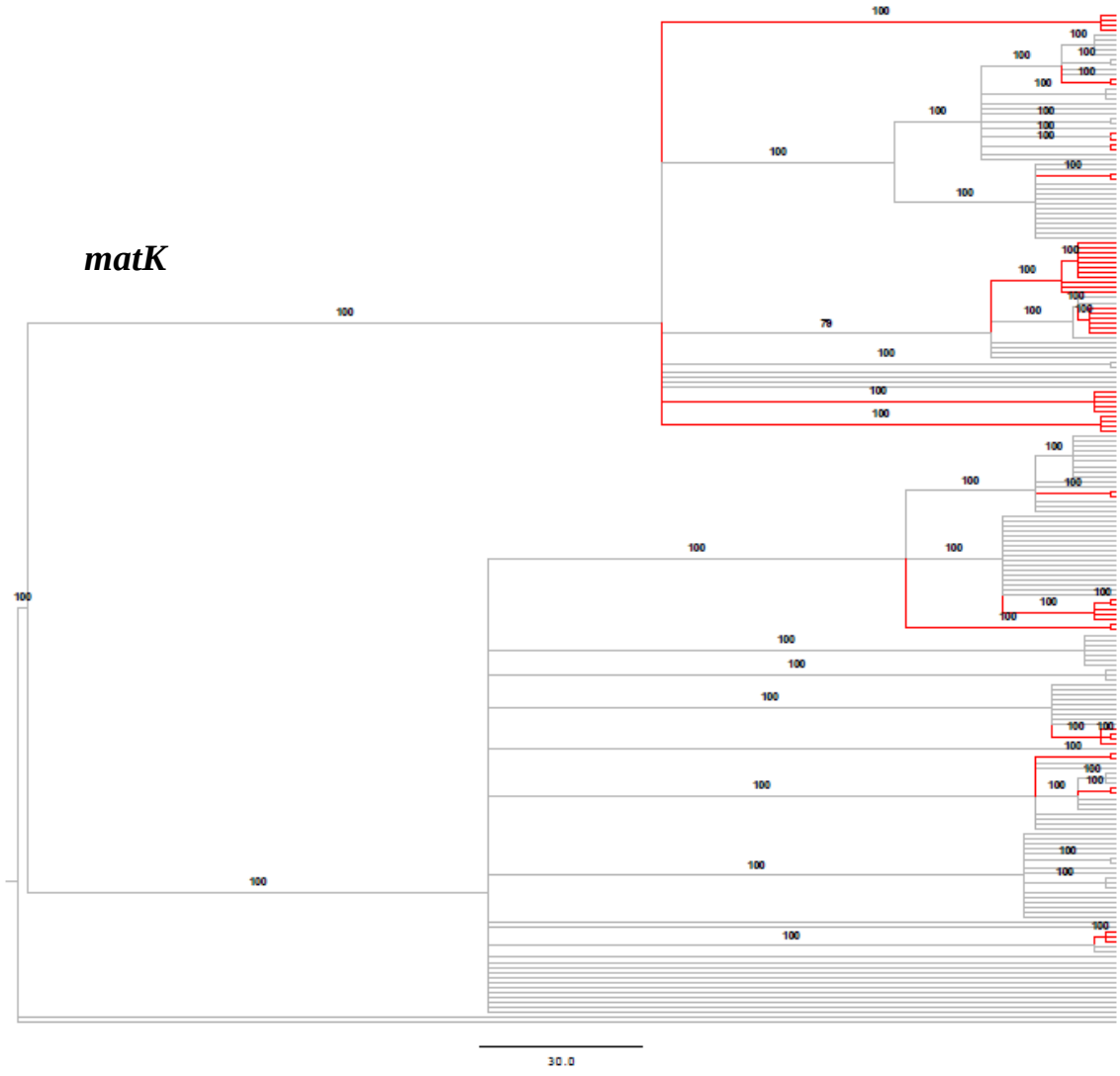

Figure 5. Continued

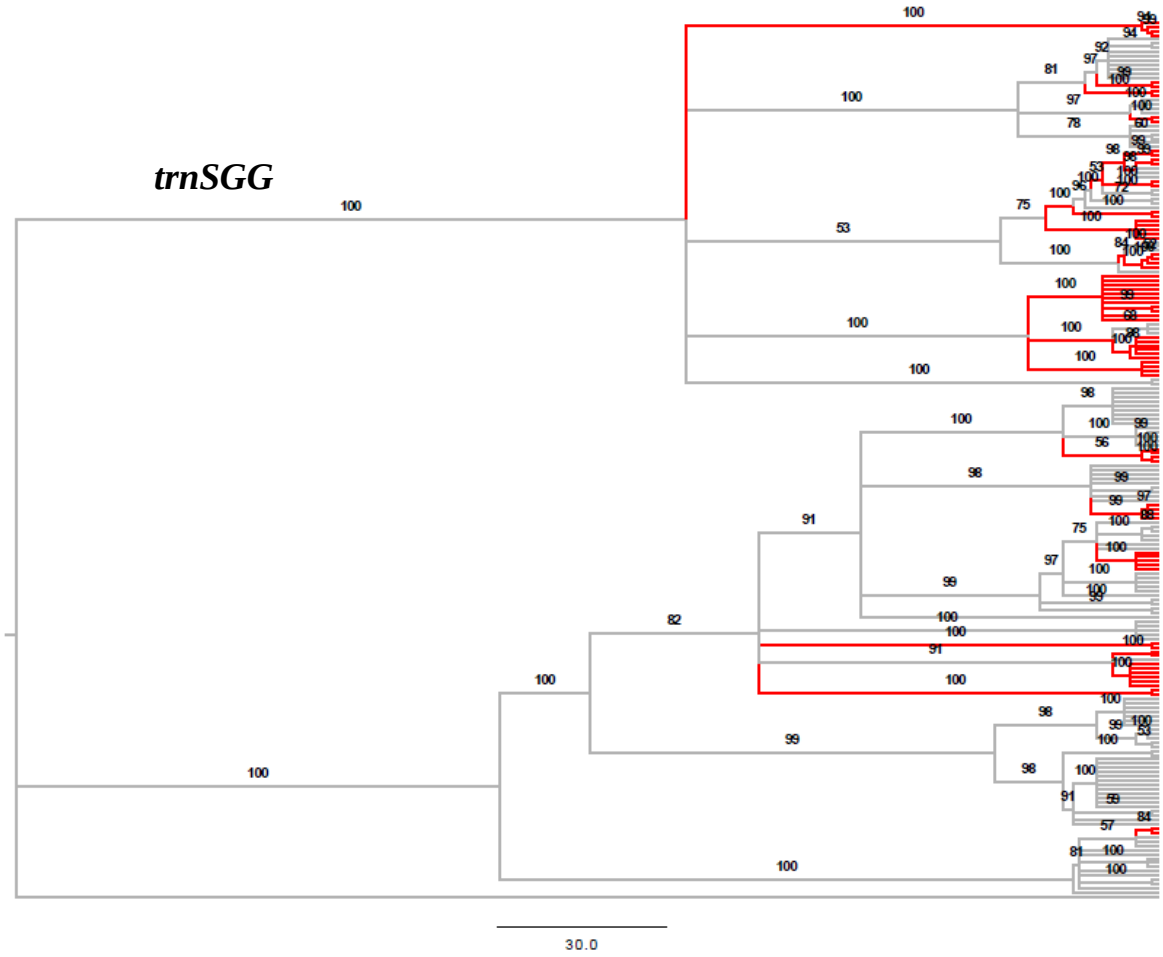

Figure 5. Continued

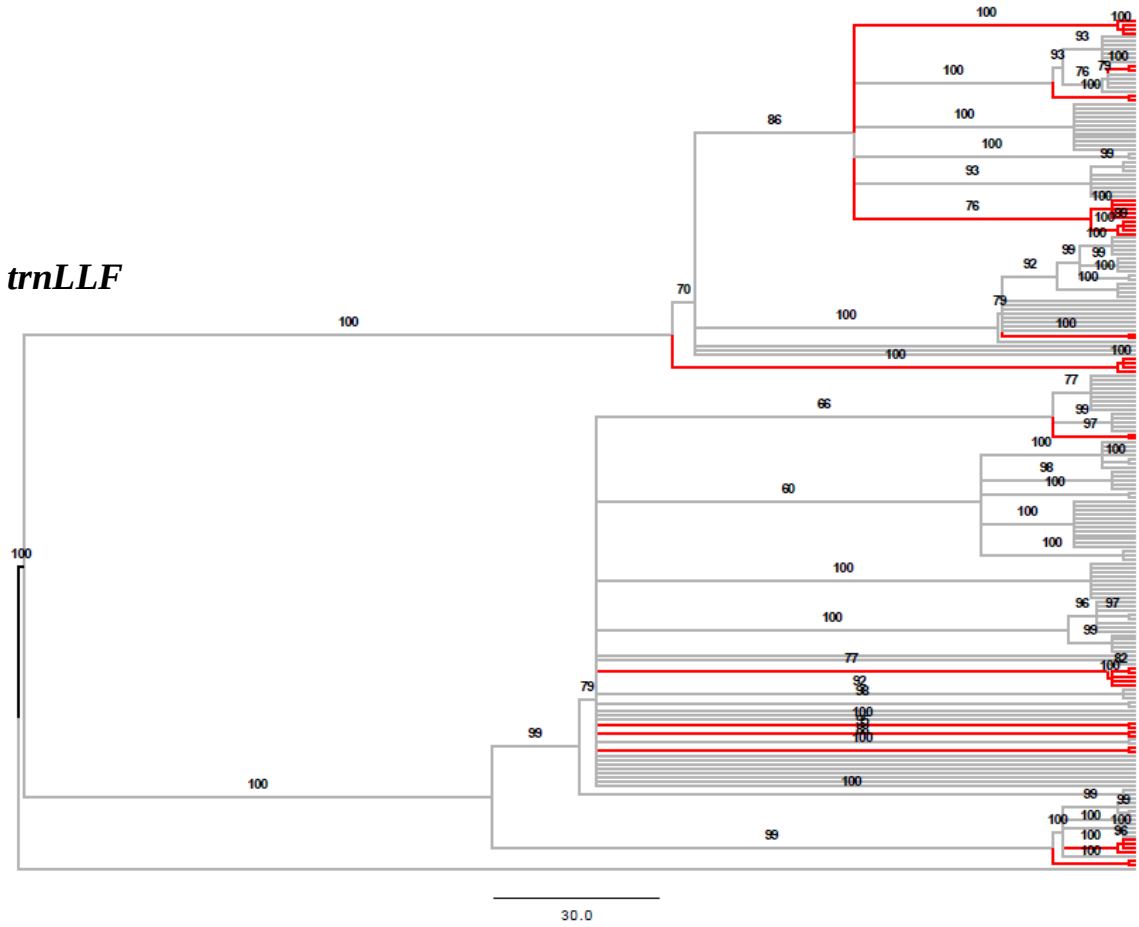

Figure 5. Continued

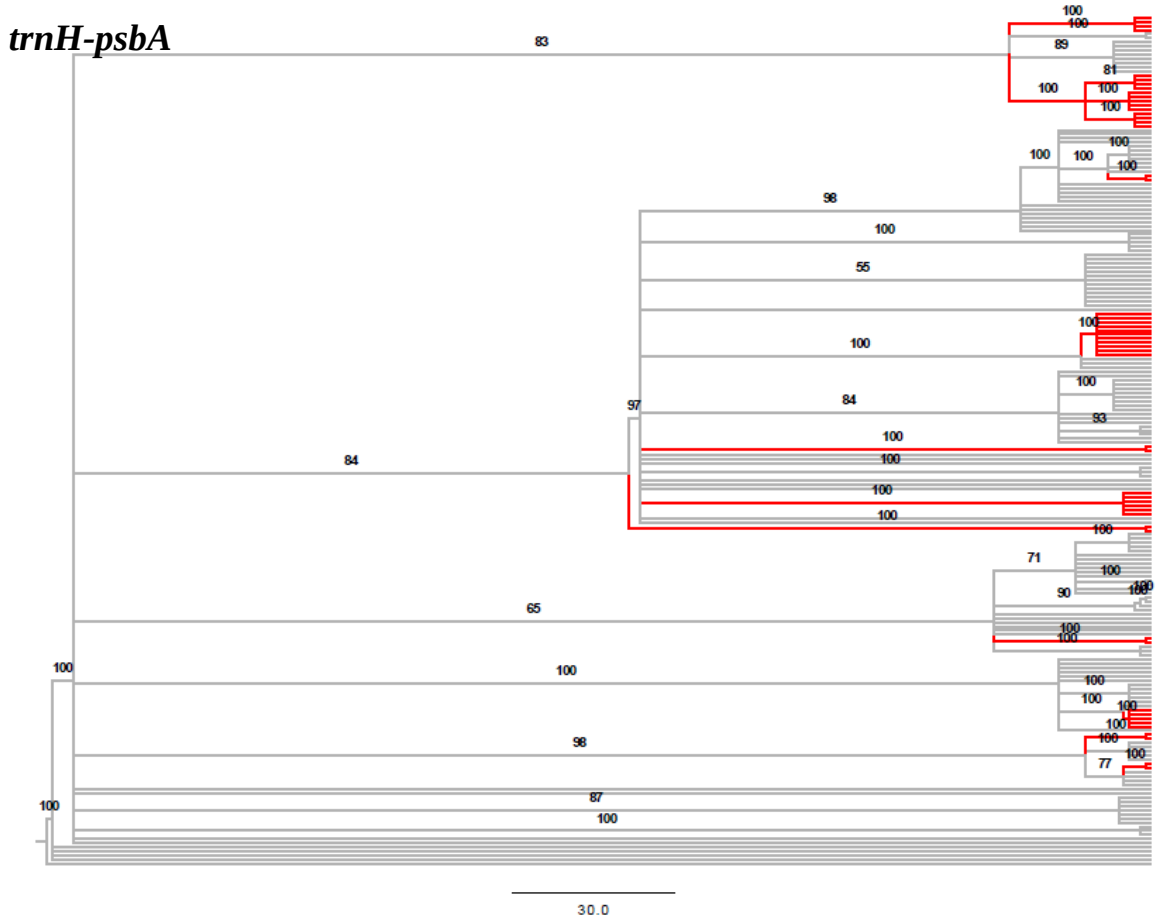

Figure 5. Continued

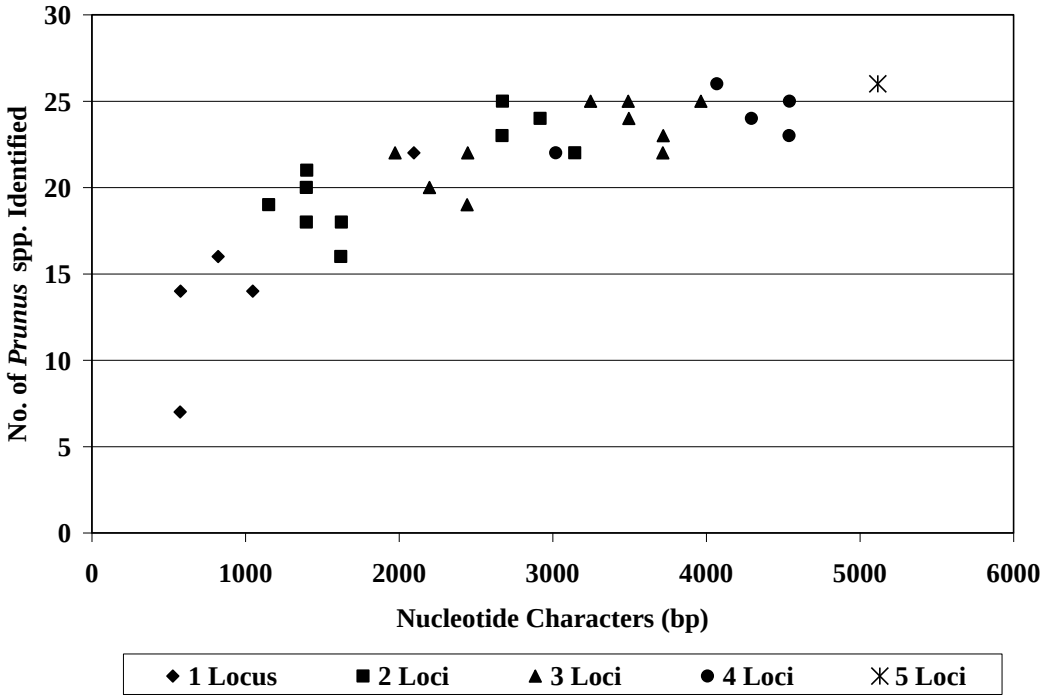

Figure 6. Relationship between sequence length and number of *Prunus* L. species identified—Graph shows that species identification begins to decelerate as the number of loci concatenated increases from two to three and hit an asymptote as the number of nucleotide characters increases to 3000 bp. Data is for 31 separate analyses using Bayesian analyses. bp = base pairs; spp.= species; No. = Number.

Literature Cited

Alfred P. Sloan Foundation. Barcode of Life. <http://www.sloan.org/program/7>. Accessed 27 August 2009.

APG II. 2003. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141: 399-436.

Asahina, H., J. Shinozaki, K. Masuda, Y. Morimitsu, and M. Satake. 2010. Identification of medicinal *Dendrobium* species by phylogenetic analyses using *matK* and *rbcL* sequences. *Journal of Natural Medicines* 64: 133-138.

Bailey, L.H. and E.Z. Bailey. 1941. Hortus second: a concise dictionary of gardening, general horticulture, and cultivated plants in North America. Macmillan, New York, New York, USA.

Blackburn, B. 1952. Trees and shrubs in eastern North America: keys to the wild and cultivated woody plants in the temperate regions exclusive of conifers. Oxford University Press, New York, New York, USA.

Blaxter, M.L., J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe. 2005. Defining operational taxonomic units using DNA barcode data. P*hilosophical Transactions of the Royal Society B* 360: 1935-1943.

Borchsenius, F. 2009. Department of Biological Sciences, University of Aarhus, Denmark. Available online at: http://192.38.46.42.aubot/fb/FastGap_home.htm

Borris, H., H. Brunke, and M. Keith. 2006. Almond profile: updated 2009 by D.Huntrods. *Agricultural Marketing Resources Center*.

Bortiri, E., S. Oh, J. Jiang, S. Baggett, A. Granger, C. Weeks, M. Buckingham, D. Potter, and D.E. Parfitt. 2001. Phylogenetic and systematics of *Prunus* (Rosaceae) as determined by sequence analysis of ITS and the chloroplast *trnL-trnF* spacer DNA. *American Society of Plant Taxonomists* 26: 797-807.

Bortiri, E., S. Oh, F. Gao, and D. Potter. 2002. The phylogenetic utility of nucleotide sequences of sorbitol 6-phosphate dehydrogenase in *Prunus* (Rosaceae). *American Journal of Botany* 89: 1697-1708.

Bortiri, E., B. Vanden Heuvel, and D. Potter. 2006. Phylogenetic analysis of morphology in *Prunus* reveals extensive homoplasy. *Plant systematics and Evolution* 259: 53-71.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA* 106: 12794-12797.

Chase, M.W., N. Salamin, M. Wilkinson, J.M. Dunwell, R.P. Kesanakurthi, N. Haidar, and V. Savolainen. 2005. Land plants and DNA barcodes: short-term and long-term goals. P*hilosophical Transactions of the Royal Society B* 360: 1889-1895.

Chase, M.W., R.S. Cowan, P.M. Hollingsworth, C. van den Berg, S. Madriñán, G. Petersen, O. Seberg, T. Jørgsensen, K.M. Cameron, M. Carine, N. Pedersen, T.A.J. Hedderson, F. Conrad, G.A. Salazar, J.E. Richardson, M.L. Hollingsworth, T.G. Barraclough, L. Kelly, and M. Wilkinson. 2007. A proposal for a standardised protocol to barcode all land plants. *Taxon* 56: 295-299.

Chen S., H. Yao, J. Han, C. Liu, J. Song, L. Shi, Y. Zhu, X. Ma, T. Gao, X. Pang, K. Luo, Y. Li, X. Li, X. Jia, Y. Lin, and C. Leon. 2010. Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. *Plos One* 5: e8613. doi:10.1371/journal.pone.0008613.

Clare, E.L., K.C.R. Kerr, T.E. von Königslöw, J.J. Wilson, and P.D.N. Hebert. 2008. Diagnosing Mitochondrial DNA diversity: applications of a sentinel gene approach. *Journal of Molecular Evolution* 66: 362-367.

Correll, D.S. and M.C. Johnston. 1970. Manual of the vascular plants of Texas. Texas Research Foundation, Renner, Texas, USA.

Cunningham, A.B. and U. Schippman. 1997. Trade in *Prunus afrciana* and the implementation of CITES. German Federal Agency for Nature Conservation.

Devey, D.S., M.W. Chase, and J.J. Clarkson. 2009. A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in noncoding plastid regions. *Taxon* 58: 7-15.

Duncan, W.H., and M.B. Duncan. 1988. Tress of the southeastern United States. University of Georgia Press, Athens, Georgia, USA.

Edwards, D., A. Horn, D. Taylor, V. Savolaninen, and J.A. Hawkins. 2008. DNA barcoding of a large genus, *Aspalathus* L. (Fabaceae). *Taxon* 57: 1317-1327.

Erickson, D.L., J. Spouge, A. Resch, L.A. Weigt, and J. Kress. 2008. DNA barcoding in land plants: developing standards to quantify and maximize success. *Taxon* 57: 1304-1316.

Farrington, L., P. MacGillivray, R. Faast, and A. Austin. 2009 Investigating DNA barcoding options for the identification of *Caladenia* (Orchidaceae) species. *Australian Journal of Botany* 57: 276-286.

Fazekas, A.J., K.S. Burgess, P.R. Kesanakurti, S.W. Graham, S.G. Newmaster, B.C. Husband, D.M. Percy, M. Hajibabaei, and S.C.H. Barrett. 2008. Multiple Multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PlosOne* 3: e2802. doi: 10.1371/journal.pone.0002802.

Fazekas, A.J., P.R. Kesanakurti, K.S. Burgess, D.M. Percy, S.W. Graham, S.C.H. Barrett, S.G. Newmaster, M. Hajibabaei, and B.C. Husband. 2009. Are plant species inherently harder to discriminate than animal species using DNA barcode markers?. *Molecular Ecology Resources* 9: 130-139.

Fazekas, A.J., R. Steeves, S.G. Newmaster, and P.M. Hollingsworth. 2010. Stopping the stutter: Improvements in sequence quality from regions with mononucleotide repeats can increase the usefulness of non-coding regions for DNA barcoding. *Taxon* 59: 694-697.

Fernald, M.L. 1950. Gray's manual of botany, 8th ed. American Book, New York, New York, USA

Flora of North America Editorial Committee, eds. 1993+. Flora of North America North of Mexico. 16+ vols. New York and Oxford.

Food and Agriculture Organization of the United Nations Website. TradeSTAT: Crops and livestock products. <http://faostat.fao.org> Accessed 27 August 2009.

Ford, C.S., K.L. Ayers, N. Toomey, N. Haider, J. Van Alphen Stahl, L.J. Kelly, N. Wikstrom, P.M. Hollingsworth, R.J. Duff, S.B. Hoot, R.S. Cowan, M.W. Chase, and D M.J. Wilkinson. 2009. Selection of candidate coding DNA barcoding regions for use on land plants. *Botanical Journal of the Linnean Society* 159: 1-11.

Gielly, L., and P. Taberlet. 1994. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus *rbcL* sequences. *Molecular Biology and Evolution* 11: 769-777.

Gene Codes. 2007. Sequencher v. 4.7. Gene Codes, Ann Arbor, Michigan, USA.

Gleason, H.A. 1952. The new Britton and Brown illustrated flora of the United States and Canada, vol. 2. Lancaster Press, Lancaster, Pennsylvania, USA.

Gleason, H.A., and A. Cronquist. 1991. Manual of vascular plants of northeastern United States and adjacent Canada, 2nd ed. New York Botanical Garden, Bronx, New York, USA.

Godfrey, R.K. 1988. Trees, shrubs, and woody vines of northeastern Florida and adjacent Georgia and Alabama. University of Georgia Press, Athens, Georgia, USA

Haider, N. 2003. Developments and use of universal primers in plants. PhD Thesis, University of Reading, United Kingdom.

Hajibabaei, M., G.A.C. Singer, P.D.N. Hebert, and D.A. Hickey. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics* 23: 167-172.

Hasagawa, M., K. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22: 160-174.

Hebert, P.D.N., A. Cywinska, S.L. Ball, and J.R. De Ward. 2003. Biological identification through DNA barcodes. *Proceedings of the Royal Society of London B* 270: 313-321.

Hebert, P.D.N., M.Y. Stoeckle, T.S. Zemlack, and C.M. Frances. 2004. Identification of birds through DNA barcodes. *PlosOne* 2: 16571663.

Hilu, K., D.T. Borsch, K. Müller, D.E. Soltis, P.S. Soltis, V. Savolainen, M.W. Chase, M.P. Powell, L.A. Alice, Evans R., H. Sauquet, C. Neinhuis, T.A.B. Slotta, J.G. Rohwer, C.S. Campbell, and L.W. Chatrou. 2003. Angiosperm phylogeny based on *matK* sequence information. *American Journal of Botany* 90: 1758–1776.

Hollingsworth, M.L., A.A. Clark, L.L. Forrest, J. Richardson, R.T. Pennington, D.G.

Long, R.S. Cowan, M.W. Chase, M. Gaudeul, and P.M. Hollingsworth. 2009. Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources* 9: 439-457.

Keane, T.M., T.J. Naughton, and J.O. McInerney. 2007. MultiPhyl: A high-throughput phylogenomics webserver using distributed computing. *Nucleic Acids Research* 35: W33-W37.

Kelchner, S. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of Missouri Botanical Garden* 87: 482-498.

Kelly, L.J., G.K. Ameka, and M.W. Chase. 2010. DNA barcoding of African Podostemaceae (river-weeds): A test of proposed barcode regions. *Taxon* 59: 251-260.

Kress, W.J., K.J. Wurdack, E.A. Zimmer, L.A. Weigt, and D.H. Janzen. 2005. Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences, USA* 102: 8369--8374.

Kress, W.J., and D.L. Erickson. 2007. A two-locus global DNA barcode for land plants: the coding rbcL gene complements the noncoding trnH-psbA spacer region. *PlosOne* 1: 1 - 10.

Kress. W.J., and D.L. Erickson. 2008. DNA barcodes: genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences, USA* 105: 2761-2762.

Kress, W.J., D.L. Erickson, F.A. Jones, N.G. Swenson, R. Perez, O. Sanjur, and E. Bermingham. 2009. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences of the, USA* 106(44): 18627-18632.

Krussman, G. 1978. Manual of cultivatedbroad-leaved trees and shrubs. Vol. Itt, Pru-Z(translated 1986). Timber Press, Portland,Ore.

Lahaye, R., M. van der Bank, D. Bogarin, J. Warner, F. Pupulin, G. Gigot, O. Maurin, S. Duthoit, T.G. Barraclough, and V. Savolainen. 2008. DNA barcoding the floras of biodiversity hotspots. *Proceeding of the National Academy of Sciences, USA* 105: 2923-2928.

Larkin M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentini, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948.

Lee, S. and J. Wen. 2001. A phylogenetic analysis of *Prunus* and the Amygdaloideae (Rosaceae) using ITS sequences of nuclear ribosomal DNA. *American Journal of Botany* 88:1150-1160.

Lee, S-B., C. Kaittanis, R.K. Jansen, J.B. Hostetler, L.J. Tallon, C.D. Town, and H. Daniell. 2006. The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. *BMC Genomics* 7: doi: 10.1186/1471-2164/7/61.

Lingdi, L., G. Cuizhi, L. Chaoluan, C. Alexander, B. Bartholomew, A.R. Brach, D.E. Boufford, H. Ikeda, H. Ohba, K.R. Robertson, and S.A. Spongberg. 2003. Rosaceae In: Z. Y. Wu, P. H. Raven, and D. Y. Hong, eds. 2003. Flora of China. Vol. 9 (Pittosporaceae through Connaraceae). Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis, USA.

Maddison, D.R., W.P. Maddison. 2001. MacClade 4.06. Sinauer, Sunderland Massachusetts.

Maddison, D.R., and K.-S. Schulz (eds.) 2007. The Tree of Life Web Project. <http://tolweb.org> Accessed 27 August.

Maddison, W.P. and D.R. Maddison. 2009. Mesquite: a modular system for evolutionary analysis. Version 2.72 Available at: http://mesquiteproject.org

Meier, R., K. Shiyang, G. Vaidya, and P.K.L Ng. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* 55: 715-728.

Meyer, C.P. and G. Paulay. 2005.DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biololgy* 3: e422

Mitchell, A. 2008. DNA barcoding demystified. *Australian Journal of Entomology* 47: 169-173.

Mowrey B. D. and D. J. Werner. 1990. Phylogenetic relationships among species of *Prunus* as inferred by isozyme markers. *Theoretical and Applied Genetics* 80: 129-133

National Clonal Germplasm Repository for fruit and nut crops, Davis, CA. < http://www.ars.usda.gov/main/site_main.htm?modecode=53-06-20-00> Accessed 27 August 2009.

Newmaster, S.G., A.J. Fazekas, and S. Ragupathy. 2006. DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Canadian Journal of Botany* 84: 335-341.

Newmaster, S.G., A.J. Fazekas, R.A.D. Steeves, and J. Janovec. 2008. Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources* 8: 480-490.

Newmaster, S.G., and S. Ragupathy. 2009. Testing plant barcoding in a sister species complex of pantropical Acacia (Mimosoideae, Fabaceae). *Molecular Ecology Resources* 9: 172–180.

Nitta, J.H. 2008. Exploring the utility of three plastid loci for biocoding the filmy ferns (Hymenophyllaceae) of Moorea. *Taxon* 57:3 725-736.

Ragupathy, S., S.G. Newmaster, M. Murugesan, and V. Balasubramaniam. 2009. DNA barcoding discriminates a new cryptic grass species revealed in an ethnobotany study by the hill tribes of the Western Ghats in southern India. *Molecular Ecology Resources* 9: 164-171.

Rieseberg, L.H. 1997. Hybrid origins of plant species. *Annual Review of Ecology and Systematics* 28: 359-389.

Palmer, J.D., K.L. Adams, Y. Cho, C.L. Parkinson, Y-L. Qiu, and K. Song. 2000. Dynamic evolution of plant mitochondrial genomes: Mobile genes and introns and highly variable mutation rates. *Proceedings of the National Academy of Sciences, USA* 97: 6960-6966.

Potter, D., T. Eriksson, R.C. Evans, S. Oh, J.E.E. Smedmark, D.R. Morgan, M. Kerr, K.R. Robertson, M. Arsenault, T.A. Dickinson, and C.S. Campbell. 2007. Phylogeny and classification of Rosaceae. *Plant Systematics and Evolution* 266: 5-43.

Radford, A.E., H.E. Ahles, and C.R. Bell. 1968. Manual of the vascular flora of the Carolinas. University of North Carolina Press, Chapel Hill, North Carolina, USA.

Rheder, A. 1940 Manual of cultivated trees and shrubs hardy in North America, exclusive of the subtropical and warmer temperate regions, 2nd revised and enlarged edition. Macmillan, New York, USA.

Ronquist, F., and J.P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572-1574.

Sass, C., D.P. Little, D.W. Stevenson, and C.D. Specht. 2007. DNA Barcoding in the Cycadales: Testing the Potential of Proposed Barcoding Markers for Species Identification of Cycads. *Plos One* 2: e1154 doi: 10.1371/journal.pone.0001154.

Seberg, O., and G. Petersen. 2009. How many loci does it take to DNA barcode a Crocus? *PlosOne* 4: e4598. doi: 10.1371/journal.pone.0004598.

Shaw, J. and R.L. Small. 2004. Addressing the "hardest puzzle in American pomology:" Phylogeny of *Prunus* section *Prunocerasus* (Rosaceae) based on seven noncoding chloroplast DNA regions. *American Journal of Botany* 91: 985-996.

Shaw, J. and R.L. Small. 2005. Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*; Rosaceae). *American Journal of Botany* 92: 2011-2030.

Shaw, J., E.B. Lickey, J.T. Beck, S.B. Farmer, W. Liu, J. Miller, K.C. Siripun, C.T. Winder, E.E. Schilling, and R.L. Small. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92: 142–166.

Shaw, J., E.B. Lickey, E.E. Schilling, and R.L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* 94: 275–288.

Small, J.K. 1933. Manual of the southeastern flora. Science Press Printing, Lancaster, Pennsylvania, USA.

Smith, E.B. 1994. Keys to the flora of Arkansas. University of Arkansas Press, Fayetteville, Arkansas, USA.

Spooner, D.M. 2009. DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. *American Journal of Botany* 96: 1177-1189.

Starr, J.R., R.F.C. Naczi, and B.N. Chouinard. Plant DNA barcodes and species resolution in sedges (*Carex*, Cyperaceae). *Molecular Ecology Resources* 9: 151-163.

Steele, P.R., L.M. Friar, L.E. Gilbert, and R.K. Jansen. 2010. Molecular systematics of the neotropical genus *Psiguria* (Cucurbitaceae): Implications for phylogeny and species identification. *American Journal of Botany*. 97: 156-173.

Stewart, K.M. 2003. The African Cherry (*Prunus Africana*): from hoe-handles to the international herb market. *Economic Botany* 57: 559-569.

Steyermark, J.A. 1963. Flora of Missouri. Iowa State University Press, Ames, Iowa, USA.

Swofford, D.L. 2001. PAUP 4.0: Phylogenetic analysis using parsimony (and other methods). Sinauer Associates, Sunderland, MA.

Taberlet, P., E. Coissac, F. Pompanon, L. Gielly, C. Miquel, A. Valentini, T. Veramt, G. Corthier, C. Brochmann, and E. Willerslev. 2006. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research* 35: e14.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57-86.

Tautz, D., P. Arctander, A. Minelli, R.H. Thomas, and A.P. Vogler. 2003. A plea for DNA taxonomy. *Trends in Ecology and Evolution* 18: 70-74.

Valentini, A., F. Pompanon, and P. Taberlet. 2009. DNA barcoding for ecologists. *Trends in Ecology and Evolution* 24:2 110-117.

Ward, R.D., T.S. Zemlak, B.H. Innes, P.R. Last, and P.D.N. Hebert. 2005. DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B* 360: 1847-1857.

Wen, J., S.T. Berggren, C. Lee, S. Ickert-Bond, T. Yi, K. Yoo, L. Xie, J. Shaw, and D. Potter. 2008. Phylogenetic inferences in *Prunus* (Rosaceae) using chloroplast *ndhF* and nuclear ribosomal ITS sequences. *Journal of Systematics and Evolution* 46: 322-332.

Will, K.W., and D. Rubinoff. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20: 47-55.

Wilson, E.O. 2000. A global map of biodiversity. *Science*. 289:5488 2279.

Wofford, B.E., and E.W. Chester. 2002. Guide to the trees, shrubs, and woody vines of Tennessee. University of Tennessee Press, Knoxville, Tennessee, USA.

Wolfe, K.H., W-H. Li, and P.M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* 84: 9054-9058.

Wunderlin, R.P. 1988. Guide to the vascular plants of Florida. University of Florida Press, Gainesville, Florida, USA.

Zhang, W., J.F. Wendel, and L.G. Clarke. 1997. Bamboozled again! Inadvertent isolation of fungal rDNA sequences from Bamboos (Poaceae: Bambusoideae). *Molecular Phylogenetics and Evolution* 8: 205-217.

Ziegenhagen, B., B. Fady, V. Kuhlenkamp, and S. Liepelt. 2005. Differentiating groups of *Abies* species with a simple molecular marker. *Silvae Genetica* 54: 123-126.

CHAPTER 2

FUTURE DIRECTIONS OF DNA BARCODING

Introduction

DNA barcoding is a burgeoning field that has the potential to help better identify species and better understand evolutionary dynamics at the species level. The PWG proposed in late 2009 that two coding chloroplast regions, *rbcL* and *matK*, serve as the barcodes for all plants with the goal of standardizing and unifying the plant barcoding community. Despite the call for further evaluation of *matK* and *rbcL* by CBOL PWG (2009), BOLD has begun accepting *matK*+*rbcL* data for plant barcoding. My thesis research has shown that *matK*+*rbcL* contain low levels of genetic variability when compared to 34 noncoding cpDNA regions that were previously tested by Shaw et al. (2005; 2007). I have also shown that this combination is a poor species barcode for *Prunus* L. A direct test of species identification showed the < 50% of *Prunus* species could be identified using *matK*+*rbcL*. My data show that there are other cpDNA regions that should be tested before the plant barcoding community settles on these two regions.

Identifying a particular region(s) is not the only concern facing the plant barcoding community. Within the constructs of DNA barcoding, issues surrounding species sampling and analytics need to be evaluated by the barcoding community in order to improve species identification. These key aspects should be improved before the barcoding community goes any further. Below is a brief discussion on why these two issues are important to the future of DNA barcoding as a whole and suggestions for future projects.

Discussion

Species Sampling

Species sampling has long been a problem in molecular phylogenetic studies, due to the inherent costs associated with generating sequence data for each specimen, but costs continue to decrease and improved automated sequencing has decreased the time to obtain sequence data. However, species identification using DNA barcodes relies on measuring the amount of genetic variability within and between species. To date, there is no consensus as to how many specimens are needed to build a species identification database. Taberlet et al. (2006) suggested that at least ten individuals per species be sequenced to capture genetic variation within a species, while Matz and Nielsen (2005) proposed 12 individual per species. To resolve this problem, researchers should sample heavily within populations to adequately capture the amount of genetic diversity within a species before any database should be considered reliable. Currently there are just over 88,000 formally described species with a barcode (see BOLD: http://www.boldsystems.org/views/login.php) and most species in BOLD are represented by ≤ 10 sequences (Hajibabaei et al., 2007) and the chances of assigning an unknown sample to species could be great since intra- and interspecific genetic variation may not been adequately captured in the database.

Recently, Zhang et al. (2010) looked more closely at how many samples per animal species are needed to sufficiently measure genetic variability within a species. Using *COI* sequence data from a species of butterflies (data obtained from original study by Hebert et al., 2004) and simulated data, they found that sampling 5-10 individuals per species was far from adequate to capture most genetic diversity within a species and further argue that most investigations have under sampled by at least 400% (Zhang et al., 2010). More importantly,

Zhang et al., (2010) note that there is no magic number for sample size, i.e. it is species-dependent, so it is up to individual researchers to determine what the appropriate sample size is for each species they want to sample. Future plant barcoding research should look at how many species are needed to capture all possible haplotypes in plants since Zhang et al. (2010) focused on animal species. Considering the unique evolutionary histories of plants it would be interesting to see if the sampling needs in plants would be markedly different than those noted by Zhang et al. (2010).

Another potential study based on Zhang et al. (2010), could look at sampling species across their ranges to measure the amount of genetic variability within a single species complex in order to correctly capture all possible haplotypes. For example, studies could look at the genetic variability contained within species found in the following three types of ranges: narrow endemic species, a broad ranging species with a semi-restricted habitat, and a species with a cosmopolitan distribution. Intraspecific and interspecific variation could be markedly different for these three species depending on range overlap with sister species, since spatial closeness promotes haplotype sharing between closely related species (Shaw et al., 2004; 2005).One also has to consider time since divergence from most common recent ancestor, since reproductive barriers may not be fully developed allowing for hybridization. Following a sampling scheme like this, could help researchers better estimate how many samples per species are needed in order to capture genetic diversity within a species.

Until species sampling requirements are better understood, the assignment of unknown specimens to species will continue to be unreliable. This endeavor will not come without increased costs but it is imperative to building a better infrastructure. Ironically, increased sampling means that DNA barcoding bioinformatics and search algorithms, used by GenBank

and BOLD, are going to have to be significantly improved. Add to the fact that plant barcoding is going to require more than a single locus and this could be a difficult bioinformatics challenge.

Analytics

Analysis of species identification is a burgeoning field that is adapting quickly to the unique needs of DNA barcoding analytics (see Sarkar, 2009). An understanding of DNA barcoding bioinformatics is one of the most pressing needs for the barcoding community since it is the backbone of species identification. There are currently no agreed upon analytical protocols in place, so it is difficult to determine which markers will identify the most species since differing analytical methods may produce different results, as I show in the preceding part.

Currently, CBOL recommends the use of uncorrected pairwise distances for species identification; however, my research found tree building analyses to be slightly more robust than distance based metrics in some gene region combinations. Unfortunately, tree building analyses are slower to run and require the use of multiple samples per species to accurately depict relationships, so they may not be as useful as distance metrics are for search databases such as BOLD and GenBank. Distance metrics, on the other, rely on selecting an arbitrary threshold at which point one considers two samples separate species. Right now, that threshold has been set at 3% but there is no empirical data to suggest that this value has any validity. Future endeavors should test whether or not the 3% threshold is valid for species identification purposes.

Erickson et al. (2008) argue that bioinformatics should influence the selection of a plant barcode since analysis of data places requirements on the feasibility of database and algorithm design. They further point out that noncoding gene regions present a challenge to database design since they increase search times due to the frequent presence of indel characters (Erickson et al. 2008). Following this assertion, future projects could utilize the sequence data generated

for the 256 *Prunus* L. samples to test search times on BOLD or GenBank's BLAST for coding and noncoding gene regions to see if there is a significant difference between submitting a query and obtaining results, as well as how often the database returns a correct or incorrect identification. Future projects could also focus on new species identification programs (as they become available) and compare the results from this study to determine if these new programs improve species identification.

With the speed at which barcoding bioinformatics is changing it is going to be difficult to unify the field in the short term. Developing search databases and the analytics behind them requires that species sampling be adequate to capture all possible haplotypes as well as understanding evolutionary rates of change at the DNA level. Without these pieces the chances of assigning a sample to species incorrectly increases. Simply put, the DNA barcode library is still being constructed, and it is clear that the analytics will continue to improve species identification results.

Conclusions

DNA barcoding continues to garner more and more attention, but it is clear that there is a need to establish more rigorous standards and improve upon current ones. My research suggests that there are other cpDNA regions that may be more suitable barcodes than *matK* and *rbcL* and that the search for a plant barcode should not be over, unless we are satisfied with identifying 70% of plants. There are a number of other areas that the barcoding community must still address, most notably species sampling and bioinformatics.

The barcoding community needs to develop sampling standards and strategies as the field moves forward in order to ensure that within and between species genetic variation is accurately accounted. Without these improved sampling efforts it will be difficult to improve species

identification bioinformatics or search databases that are used to identify unknown specimens. Improvement to species sampling will undoubtedly enhance and improve the quality of queries from search databases.

The unique evolutionary histories of plants pose many hurdles for the plant barcoding to overcome, most notably incomplete lineage sorting, chloroplast sharing between closely related species, and the lack of genetic divergence between the chloroplasts of closely related species. There is still a lot of work to be done before DNA barcoding can be considered as successful in plants as it is in animals, generally speaking, but success is not inevitable. Despite the obstacles, DNA barcoding of the world's flora continues to move forward.

Literature Cited

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceeding of the National Academy of Sciences* 106:31 12794-12797.

Erickson, D.L., J. Spouge, A. Resch, L.A. Weigt, and J. Kress. 2008. DNA barcoding in land plants: developing standards to quantify and maximize success. *Taxon* 57:4 1304-1316.

Hajibabaei, M., G.A. Singer, E.L. Clare, P.D.N. Hebert. 2007. Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology* 5: 24.

Hebert, P.D.N., E.H. Penton, J.M. Burns, D.H. Janzen, and W. Hallwachs. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences, USA* 101: 14812–14817.

Matz, M.V., and R. Nielsen. 2005. A likelihood ratio test for species membership based on DNA sequence data. *Philosophical Transactions of the Royal Society B* 360: 1969–1974.

Sarkar, I.N. 2009. Biodiversity Informatics: the emergence of a field. *BMC Bioinformatics*. 10: S1

Shaw, J. and R.L. Small. 2004. Addressing the "hardest puzzle in American pomology:" Phylogeny of *Prunus* section *Prunocerasus* (Rosaceae) based on seven noncoding chloroplast DNA regions. *American Journal of Botany* 91: 985-996.

Shaw, J. and R.L. Small. 2005. Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*; Rosaceae). *American Journal of Botany* 92: 2011-2030.

Shaw, J., E.B. Lickey, J.T. Beck, S.B. Farmer, W. Liu, J. Miller, K.C. Siripun, C.T. Winder, E.E. Schilling, and R.L. Small. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92: 142–166.

Shaw, J., E.B. Lickey, E.E. Schilling, and R.L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *American Journal of Botany* 94: 275–288.

Taberlet, P., E. Coissac, F. Pompanon, L. Gielly, C. Miquel, A. Valentini, T. Veramt, G. Corthier, C. Brochmann, and E. Willerslev. 2006. Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research* 35: e14.

Zhang, A.B., L.J. He, R.H. Crozier, C. Muster, and C.-D. Zhu. 2010. Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and Evolution* 54:2010 1035-1039.

APPENDIX A

GENBANK ACCESSION NUMBERS FOR SEVEN ANGIOSPERM LINEAGES USED IN THIS STUDY

Table 7. Lineages from Shaw et al. (2005; 2007) used in this study, cpDNA regions tested and GenBank Accession Numbers. For voucher locations see Shaw et al. (2005). OG = Outgroup

Species	Source Number	*matK*	*rbcL*
Hibiscus macrophyllus Roxb.	OG6-8	HQ235319	HQ235601
Hibiscus cannabinus L.	43	HQ235320	HQ235602
Hibiscus mechowii Garcke	46	HQ235321	HQ235603
Liriodendron tulipifera L.	OG076	HQ235322	HQ235604
Magnolia acuminate L.	75	HQ235323	HQ235605
Magnolia tripetala L.	74	HQ235324	HQ235606
Minuartia uniflora (Walt.) Mattf.	OGSM3	HQ235327	HQ235607
Minuartia glabra (Michx.) Mattf.	LB2	HQ235325	HQ235608
Minuartia cumberlandensis (B.E. Wofford & Kral) McNeill	CW3	HQ235326	HQ235609
Taxodium distichum (Nutt.) Croom	IMB39	Sequence Data Not Obtained	HQ235610
Glyptostrobus pensilis (Staunton) K. Koch	240	Sequence Data Not Obtained	HQ235611
Cryptomeria japonica (L.f.) Don.	OG253	Sequence Data Obtained/ Not submitted to GenBank	HQ235612
Gratiola neglecta Torr.	OG1	HQ235328	HQ235613
Gratiola virginiana L.	005	HQ235329	HQ235614
Gratiola brevifolia Raf.	002	HQ235330	HQ235615
Pseudotrillium rivale (S. Wats.) S.B. Farmer	OG792	HQ235331	HQ235616
Trillium texanum Buckl.	794	HQ235332	HQ235617
Trillium ovatum Pursh.	779	HQ235333	HQ235618
Solanum americanum Mill.	427	HQ235334	HQ235619
Solanum ptycanthum Dunal	455	HQ235335	HQ235620
Solanum physalifolium Rusby	OG485	HQ235336	HQ235621
Carphephorus corymbosus (Nutt.) Torr. & A. Gray	652	HQ235337	HQ235622
Trilisa paniculata (Willd.) Cass.	691	HQ235338	HQ235623
Eupatorium capillifolium (Lamarck) Small	OG869	HQ235340	HQ235624
Eupatoriuum hyssopifolium L.	870	HQ235341	HQ235625
Eupatorium rotundifolium L.	002	HQ235339	HQ235626
Prunus nigra Ait.	040	HQ235342	HQ235488
Prunus virginiana L.	OG 045	HQ235343	HQ235629
Prunus hortulana Bailey	008	HQ235344	HQ235448

APPENDIX B

GENETIC VARIABILITY DATA FOR MATK AND RBCL ACROSS SEVEN ANGIOSPERM LINEAGES USED IN THIS STUDY

Table 8. Quantitative data collected for *matK* and *rbcL* in this study (see Shaw et al., 2005; 2007 for raw data for the 34 noncoding cpDNA regions). Each cell (cpDNA region/three-species survey) contains data regarding: the aligned length of the three-species surveyed; the number of indels (between the ingroup taxa and the ingroup and the outgroup taxon); the number of nucleotide substitutions (between the ingroup taxa and the ingroup and the outgroup taxon); PICs = total indels + nucleotide substitutions + inversions; the normalized PIC value; and the percent variability. Appendix abbreviations: L. = aligned length of three-species group, Subst. = nucleotide substitutions, PIC = potentially informative character. bp = base pairs.

	Metrics	***rbcL***		***matK***	
MAGNOLIID: *Liriodendron / Magnolia*	Aligned L. (bp)	*577*		*834*	
	Indels: in / out	*0*	*0*	*0*	*0*
	Subst: in / out	*0*	*16*	*4*	*9*
	Total PICs	*16*		*13*	
	Normalized PIC's	*1.524*		*1.238*	
	% variability	*2.77*		*1.56*	
MONOCOT: *Pseudotrillium / Trillium*	Aligned L. (bp)	*576*		*823*	
	Indels: in / out	*0*	*0*	*1*	*0*
	Subst: in / out	*1*	*3*	*3*	*11*
	Total PICs	*4*		*15*	
	Normalized PIC's	*0.447*		*1.676*	
	% variability	*0.69*		*1.82*	
CARYOPHYLLID: *Minuartia*	Aligned L. (bp)	*573*		*853*	
	Indels: in / out	*0*	*0*	*1*	*0*
	Subst: in / out	*0*	*7*	*9*	*35*
	Total PICs	*7*		*45*	
	Normalized PIC's	*0.458*		*2.941*	
	% variability	*1.22*		*5.28*	
EUROSID I: *Prunus*	Aligned L. (bp)	*575*		*860*	
	Indels: in / out	*0*	*0*	*2*	*3*
	Subst: in / out	*0*	*6*	*3*	*10*
	Total PICs	*6*		*18*	
	Normalized PIC's	*0.712*		*2.135*	
	% variability	*1.04*		*2.09*	

Table 8 Continued

EUROSID II: ***Hibiscus***	Aligned L. (bp)	*575*		*853*	
	Indels: in / out	*0*	*0*	*2*	*4*
	Subst: in / out	*1*	*6*	*8*	*8*
	Total PICs	*7*		*22*	
	Normalized PIC's	*0.760*		*2.389*	
	% variability	*1.22*		*2.58*	
EUASTERID I: ***Gratiola***	Aligned L. (bp)	*589*		*862*	
	Indels: in / out	*1*	*1*	*5*	*2*
	Subst: in / out	*9*	*12*	*22*	*13*
	Total PICs	*23*		*42*	
	Normalized PIC's	*1.218*		*2.225*	
	% variability	*3.90*		*4.87*	
EUASTERID II: ***Eupatorium / Carphephorus / Trilisa***	Aligned L. (bp)	*575*		*846*	
	Indels: in / out	*0*	*0*	*2*	*2*
	Subst: in / out	*1*	*6*	*6*	*12*
	Total PICs	*7*		*22*	
	Normalized PIC's	*1.049*		*3.298*	
	% variability	*1.22*		*2.60*	

APPENDIX C

PRUNUS L. ACCESSIONS USED IN THIS STUDY

Table 9. Taxa used in assessment of species identification analyses. cpDNA regions tested, source, voucher and GenBank accession numbers provided. Source and voucher information could not be found at this time for a handful of samples and have been left blank. In some cases, pieces of source and voucher information were known and have been placed in the table but bolded to show that they are not complete.

Species	**Source & Voucher**	***psbA-trnH***	***trnL Intron***	***trnL-trnF***	***trnS-trnG-trnG***	***matK***	***rbcL***
Maddenia hypoleuca396 Koehne	Wen8062	HQ188700	HQ243708	HQ243945	HQ244182	HQ235063	HQ235345
Physocarpus opulifolius084 (L.) Maxim.	JSh1015; TENN	AY500637	HQ243944	HQ244181	AY500742	HQ235318	HQ235600
Prunus africana202 (Hook. f) Kalkman	T. Eriksson 1010	HQ188701	HQ243709	HQ243946	HQ244183	HQ235064	HQ235346
Prunus africana217	Q. Luke 11470	HQ188702	HQ243710	HQ243947	HQ244184	HQ235065	HQ235347
Prunus africana260	DPRU 2557.1	HQ188703	HQ243711	HQ243948	HQ244185	HQ235066	HQ235348
Prunus africana390	DPRU 2557.2	HQ188704	HQ243712	HQ243949	HQ244186	HQ235067	HQ235349
Prunus alleghaniensis Porter var. *davisii005* Wight	G. Schmidt; MI: TENN	AY500606	AY500754,	AY500773	AY500711	HQ235072	HQ235351
Prunus alleghaniensis001 Porter	JSh834	AY500607	AY500755,	AY500774	AY500717	HQ235069	HQ235352
Prunus alleghaniensis006	JSh837	HQ188706	AY500755	AY500774	AY500712	HQ235070	HQ235353
Prunus alleghaniensis373		HQ188707	HQ243715	HQ243952	HQ244189	HQ235071	HQ235354
Prunus americana Marshall *var. lanata047* Sudw.	J. Beck 49955; TN: TENN	AY500596	AY500744	AY500763	AY500701	HQ235074	HQ235355
Prunus americana038 Marshall	JSh038, TN: TENN	AY500595	AY500743	AY500762	AY500700	HQ235073	HQ235356
Prunus amplifolia213 Pilg.	PPM-3489; Santa Cruz, Bolivia	HQ188708	HQ243716	HQ243953	HQ244190	HQ235075	HQ235357
Prunus amygdalus073 Batsch	DPRU 1463.5: TENN	AY500625	HQ243717	HQ243954	AY500730	HQ235076	HQ235358
Prunus amygdalus277	DPRU 2330.5	HQ188709	HQ243718	HQ243955	HQ244191	HQ235077	HQ235359
Prunus amygdalus283	DPRU 2434	HQ188710	HQ243719	HQ243956	HQ244192	HQ235078	HQ235360
Prunus amygdalus292	DPRU 2406.14	HQ188711	HQ243720	HQ243957	HQ244193	HQ235079	HQ235361
Prunus amygdalus296	DPRU 1128	HQ188712	HQ243721	HQ243958	HQ244194	HQ235080	HQ235362
Prunus amygdalus297	DPRU 1463.4	HQ188713	HQ243722	HQ243959	HQ244195	HQ235081	HQ235363

Table 9 Continued

Prunus amygdalus298	DPRU 1486.1	HQ188714	HQ243723	HQ243960	HQ244196	HQ235082	HQ235364
Prunus amygdalus302	DPRU 1461.1	HQ188715	HQ243724	HQ243961	HQ244197	HQ235083	HQ235365
Prunus andersonii2032 A.Gray	Wen 10630	HQ188716	HQ243725	HQ243962	HQ244198	HQ235084	HQ235366
Prunus angustifolia012 Marshall	JSh785; GA: TENN	AY500601	AY500749	AY500768	AY500706	HQ235085	HQ235367
Prunus angustifolia371		HQ188717	HQ243726	HQ243963	HQ244199	HQ235086	HQ235368
Prunus arborea (Bl.) Kalkm. var. *montana2029* (Hook.f.) Kalkman	Wen 11028	HQ188727	HQ243736	HQ243973	HQ244209	HQ235096	HQ235378
Prunus arborea (Bl.) Kalkm. var. *stipulacea2030* (King) Kalkman	Wen 11060	HQ188728	HQ243737	HQ243974	HQ244210	HQ235097	HQ235379
Prunus arborea2005 (Bl.) Kalkm.	Potter 081124-03	HQ188718	HQ243727	HQ243964	HQ244200	HQ235087	HQ235369
Prunus arborea2009	Potter 081124-05	HQ188719	HQ243728	HQ243965	HQ244201	HQ235088	HQ235370
Prunus arborea2011	Potter 081124-02	HQ188720	HQ243729	HQ243966	HQ244202	HQ235089	HQ235371
Prunus arborea2014	Potter 081124-01	HQ188721	HQ243730	HQ243967	HQ244203	HQ235090	HQ235372
Prunus arborea2015	Potter 0811233-05	HQ188722	HQ243731	HQ243968	HQ244204	HQ235091	HQ235373
Prunus arborea2028	Wen 10944	HQ188723	HQ243732	HQ243969	HQ244205	HQ235092	HQ235374
Prunus arborea351	Wen 8431	HQ188724	HQ243733	HQ243970	HQ244206	HQ235093	HQ235375
Prunus arborea402		HQ188725	HQ243734	HQ243971	HQ244207	HQ235094	HQ235376
Prunus arborea438		HQ188726	HQ243735	HQ243972	HQ244208	HQ235095	HQ235377
Prunus argentea262	DPRU 194	HQ188729	HQ243738	HQ243975	HQ244211	HQ235098	HQ235380
Prunus armeniaca L. var. *mandshurica388* Maxim.	DPRU 2311.1: TENN	AY500619	HQ243749	HQ243986	AY500724	HQ235109	HQ235391
Prunus armeniaca L. var. *shirpaivan387*	DPRU 343; RCH208	HQ188739	HQ243750	HQ243987	HQ244221	HQ235110	HQ235392
Prunus armeniaca065 L.	DPRU 1372.2; TENN	AY500620	HQ243739	HQ243976	AY500725	HQ235099	HQ235381
Prunus armeniaca246	DPRU 1685	HQ188730	HQ243740	HQ243977	HQ244212	HQ235100	HQ235382
Prunus armeniaca267	DPRU 1855.2	HQ188731	HQ243741	HQ243978	HQ244213	HQ235101	HQ235383
Prunus armeniaca269	DPRU 1794.3	HQ188732	HQ243742	HQ243979	HQ244214	HQ235102	HQ235384
Prunus armeniaca279	DPRU 2285	HQ188733	HQ243743	HQ243980	HQ244215	HQ235103	HQ235385
Prunus armeniaca282	DPRU 0692	HQ188734	HQ243744	HQ243981	HQ244216	HQ235104	HQ235386
Prunus armeniaca303	DPRU 1787.3	HQ188735	HQ243745	HQ243982	HQ244217	HQ235105	HQ235387

Table 9 Continued

Prunus armeniaca310	DPRU 1268	HQ188736	HQ243746	HQ243983	HQ244218	HQ235106	HQ235388
Prunus armeniaca312	DPRU 1754	HQ188737	HQ243747	HQ243984	HQ244219	HQ235107	HQ235389
Prunus armeniaca415	Wen 8012	HQ188738	HQ243748	HQ243985	HQ244220	HQ235108	HQ235390
Prunus avium307 L.	DPRU 1539	HQ188740	HQ243751	HQ243988	HQ244222	HQ235111	HQ235393
Prunus avium309	DPRU 1953	HQ188741	HQ243752	HQ243989	HQ244223	HQ235112	HQ235394
Prunus bifrons256 Fritsch	DPRU 1213.1	HQ188742	HQ243753	HQ243990	HQ244224	HQ235113	HQ235395
Prunus bokhariensis386 Royle ex Schneid.	DPRU 823; RCH201	HQ188743	HQ243754	HQ243991	HQ244225	HQ235114	HQ235396
Prunus brigantina266 Vill.	DPRU 937	HQ188744	HQ243755	HQ243992	HQ244226	HQ235115	HQ235397
Prunus brittoniana212 Rusby		HQ188745	HQ243756	HQ243993	HQ244227	HQ235116	HQ235398
Prunus brittoniana406	Nee & Wen 53936	HQ188746	HQ243757	HQ243994	HQ244228	HQ235117	HQ235399
Prunus bucharica366 (Korsh.) Hand.-Mazz.	DPRU 192.3	HQ188747	HQ243758	HQ243995	HQ244229	HQ235118	HQ235400
Prunus buergeriana431	Wen9356	HQ188748	HQ243759	HQ243996	HQ244230	HQ235119	HQ235401
Prunus caroliniana 048 Aiton	E. Lickeyy; FL: TENN	AY500636	HQ243760	HQ243997	AY500741	HQ235120	HQ235402
Prunus caroliniana241	JSh XXX	HQ188749	HQ243761	HQ243998	HQ244231	HQ235121	HQ235403
Prunus cerasifera076 Ehrh.	DPRU 563: TENN	AY500616	HQ243762	HQ243999	AY500721	HQ235122	HQ235406
Prunus cerasifera305	DPRU 2455	HQ188750	HQ243763	HQ244000	HQ244232	HQ235123	HQ235407
Prunus cerasifera311	DPRU 2314.1	HQ188751	HQ243764	HQ244001	HQ244233	HQ235124	HQ235408
Prunus cerasifera313	DPRU 2459.2	HQ188752	HQ243765	HQ244002	HQ244234	HQ235125	HQ235409
Prunus cerasoides2023 D. Don	Wen 10833	HQ188753	HQ243766	HQ244003	HQ244235	HQ235126	HQ235410
Prunus cerasoides434	Wen 10126	HQ188754	HQ243767	HQ244004	HQ244236	HQ235127	HQ235411
*Prunus cerasus273*L.	DPRU 1709	HQ188755	HQ243768	HQ244005	HQ244237	HQ235128	HQ235412
Prunus cerasus287	DPRU 0010	HQ188756	HQ243769	HQ244006	HQ244238	HQ235129	HQ235413
Prunus cerasus288	DPRU 1707	HQ188757	HQ243770	HQ244007	HQ244239	HQ235130	HQ235414
Prunus cerasus295	DPRU 2467	HQ188758	HQ243771	HQ244008	HQ244240	HQ235131	HQ235415
Prunus cerasus315	DPRU 1	HQ188759	HQ243772	HQ244009	HQ244241	HQ235132	HQ235416
Prunus ceylanica2019 (Wigt) Miq.	Wen 10801	HQ188760	HQ243773	HQ244010	HQ244242	HQ235133	HQ235417

Table 9 Continued

Prunus consociiflora367 Schneid.	RCH209	HQ188761	HQ243774	HQ244011	HQ244243	HQ235134	HQ235418
Prunus costata2039 (Hemsl.) Kalm.	Wen 10744	HQ188762	HQ243775	HQ244012	HQ244244	HQ235135	HQ235419
Prunus costata432		HQ188763	HQ243776	HQ244013	HQ244245	HQ235136	HQ235420
Prunus davidiana064 (Carr.) Franch.	DPRU 581: TENN	AY500626	HQ243777	HQ244014	AY500731	HQ235137	HQ235421
Prunus davidiana2034	Wen 10660	HQ188764	HQ243778	HQ244015	HQ244246	HQ235138	HQ235422
Prunus davidiana268	DPRU 2493.1	HQ188765	HQ243779	HQ244016	HQ244247	HQ235139	HQ235423
Prunus dielsiana420 Schneid.	Wen 9856	HQ188766	HQ243780	HQ244017	HQ244248	HQ235140	HQ235424
Prunus dolichobotrys2036 (Laut. & K.Sch.) Kalkm.	Wen 10703	HQ188767	HQ243781	HQ244018	HQ244249	HQ235141	HQ235425
*Prunus domestica078*L.	DPRU 350: TENN	AY500614	HQ243782	HQ244019	AY500719	HQ235142	HQ235426
Prunus domestica278	DPRU 1255	HQ188768	HQ243783	HQ244020	HQ244250	HQ235143	HQ235427
Prunus domestica289	DPRU 0927	HQ188769	HQ243784	HQ244021	HQ244251	HQ235144	HQ235428
Prunus domestica290	DPRU 1516	HQ188770	HQ243785	HQ244022	HQ244252	HQ235145	HQ235429
Prunus domestica291	DPRU 1630	HQ188771	HQ243786	HQ244023	HQ244253	HQ235146	HQ235430
Prunus domestica300	DPRU 1537	HQ188772	HQ243787	HQ244024	HQ244254	HQ235147	HQ235431
Prunus fasciculata068 (Torr.) A. Gray	DRPU 2033	AY500630	HQ243788	HQ244025	AY500735	HQ235148	HQ235432
Prunus ferganensis255 (Kostov&Rjabov) Kovalev & Kostov	DPRU 2495.3	HQ188773	HQ243789	HQ244026	HQ244255	HQ235149	HQ235433
Prunus ferganensis365	DPRU 24951.1; RCH210	HQ188774	HQ243790	HQ244027	HQ244256	HQ235150	HQ235434
Prunus ferganica394 Lincz.	DPRU1212 .1; RCH211	HQ188775	HQ243791	HQ244028	HQ244257	HQ235151	HQ235435
Prunus fordiana2024 Dunn	Wen 10845	HQ188776	HQ243792	HQ244029	HQ244258	HQ235152	HQ235436
Prunus gazelle-peninsulae2006 (Kan. & Hat.) Kalkm.	Potter 081120-02	HQ188777	HQ243793	HQ244030	HQ244259	HQ235153	HQ235437
Prunus geniculata021 R. M. Harper	JSh 898	AY500608	AY500756	AY500775	AY500713	HQ235154	HQ235438
Prunus glandulosa069 Thunb.	DPRU 403.1	AY500622	HQ243794	HQ244031	AY500727	HQ235155	HQ235439
Prunus gracilis027 Engelm. & A. Gray	JSh 936	AY500603	AY500751	AY500770	AY500708	HQ235156	HQ235440

Table 9 Continued

Prunus grayana355 Maxim.	Wen 1698-77B	HQ188778	HQ243795	HQ244032	HQ244260	HQ235157	HQ235441
Prunus grayana356	Wen 1191-77B	HQ188779	HQ243796	HQ244033	HQ244261	HQ235158	HQ235442
Prunus grayana424	Wen 9324	HQ188780	HQ243797	HQ244034	HQ244262	HQ235159	HQ235443
Prunus grisea350 (Blume ex Mull. Berol.) Kalkman	Wen 8262	HQ188781	HQ243798	HQ244035	HQ244263	HQ235160	HQ235444
Prunus grisea410		HQ188782	HQ243799	HQ244036	HQ244264	HQ235161	HQ235445
Prunus grisea414	Wen 8420	HQ188783	HQ243800	HQ244037	HQ244265	HQ235162	HQ235446
Prunus grisea437		HQ188784	HQ243801	HQ244038	HQ244266	HQ235163	HQ235447
Prunus henryi352 Schneid.	Wen 8463	HQ188785	HQ243802	HQ244039	HQ244267	HQ235164	HQ235448
Prunus henryi397		HQ188786	HQ243803	HQ244040	HQ244268	HQ235165	HQ235449
Prunus hortulana008 L.H. Bailey	JSh 821 or JSh 821-017	AY500600	AY500748	AY500767	AY500705	HQ235166	HQ235450
Prunus insititia079 L.	DPRU 2054	AY500613	HQ243804	HQ244041	AY500718	HQ235167	HQ235451
Prunus integrifolia203 Sarg.		HQ188787	HQ243805	HQ244042	HQ244269	HQ235168	HQ235452
Prunus integrifolia210		HQ188788	HQ243806	HQ244043	HQ244270	HQ235169	HQ235453
Prunus integrifolia407	Wen 8620	HQ188789	HQ243807	HQ244044	HQ244271	HQ235170	HQ235454
*Prunus japonica248*Thunb.	DPRU2248	HQ188790	HQ243808	HQ244045	HQ244272	HQ235171	HQ235455
Prunus javanica2007 T.&B. Miq.	Potter 081119-01	HQ188791	HQ243809	HQ244046	HQ244273	HQ235172	HQ235456
Prunus javanica2008	Potter 081120-01	HQ188792	HQ243810	HQ244047	HQ244274	HQ235173	HQ235457
Prunus javanica2012	Potter 081117-01	HQ188793	HQ243811	HQ244048	HQ244275	HQ235174	HQ235458
Prunus javanica435		HQ188794	HQ243812	HQ244049	HQ244276	HQ235175	HQ235459
Prunus kansuensis344 Rehder	Wen 8013	HQ188795	HQ243813	HQ244050	HQ244277	HQ235176	HQ235460
Prunus kansuensis379	DPRU 582; RCH213	HQ188796	HQ243814	HQ244051	HQ244278	HQ235177	HQ235461
Prunus kansuensis403		HQ188797	HQ243815	HQ244052	HQ244279	HQ235178	HQ235462
Prunus kuramica258 (Korsh.) Kitamura	DPRU 1467.4	HQ188798	HQ243816	HQ244053	HQ244280	HQ235179	HQ235463
Prunus lancilima2021	Wen 10829	HQ188799	HQ243817	HQ244054	HQ244281	HQ235180	HQ235464
Prunus laurocerasus080 L.	JSh1014	AY500635	HQ243818	HQ244055	AY500740	HQ235181	HQ235465
Prunus lusitanica416 L.	Wen 9462	HQ188800	HQ243819	HQ244056	HQ244282	HQ235182	HQ235466
Prunus maackii377 Rupr.	DPRU 2533	HQ188801	HQ243820	HQ244057	HQ244283	HQ235183	HQ235467
Prunus mahaleb035 L.	JSh966-116; TN: TENN	AY500631	AY500761	AY500780	AY500736	HQ235184	HQ235468
Prunus mahaleb245	DPRU 404	HQ188802	HQ243821	HQ244058	HQ244284	HQ235185	HQ235469

Table 9 Continued

Prunus malayana349 Kalkm.	Wen 8366	HQ188803	HQ243822	HQ244059	HQ244285	HQ235186	HQ235470
Prunus malayana405		HQ188804	HQ243823	HQ244060	HQ244286	HQ235187	HQ235471
Prunus maritima Marshall var. *gravesii002* (Small) G. J. Anderson	G.J. Anderson; TENN	AY500610	AY500758	AY500777	AY500715	HQ235189	HQ235472
Prunus maritima007 Marshall	JSh877-045; MA: TENN	AY500609	AY500757	AY500776	AY500714	HQ235188	HQ235473
Prunus mexicana028 S. Watson	JSh919; TX: TENN	AY500599	AY500747	AY500766	AY500704	HQ235190	HQ235474
Prunus mira071 Koehne	DPRU 2228.3: TENN	AY500627	HQ243824	HQ244061	AY500732	HQ235191	HQ235475
Prunus mira294	DPRU 2228.2	HQ188805	HQ243825	HQ244062	HQ244287	HQ235192	HQ235476
Prunus mira308	DPRU 2583.12	HQ188806	HQ243826	HQ244063	HQ244288	HQ235193	HQ235477
Prunus mira314	DPRU 2232	HQ188807	HQ243827	HQ244064	HQ244289	HQ235194	HQ235478
Prunus mira316	DPRU 2561.26	HQ188808	HQ243828	HQ244065	HQ244290	HQ235195	HQ235479
Prunus mira417	Wen 9171	HQ188809	HQ243829	HQ244066	HQ244291	HQ235196	HQ235480
Prunus mume066 Siebold & Zucc.	DPRU 1588: TENN	AY500621	HQ243830	HQ244067	AY500726	HQ235197	HQ235481
Prunus mume250	DPRU 2427.1	HQ188810	HQ243831	HQ244068	HQ244292	HQ235198	HQ235482
Prunus mume306	DPRU 2426.1	HQ188811	HQ243832	HQ244069	HQ244293	HQ235199	HQ235483
Prunus mume430	Wen9182	HQ188812	HQ243833	HQ244070	HQ244294	HQ235200	HQ235484
Prunus munsoniana013 W.Wight & U.P. Hedrick	JSh 810	AY500602	AY500750	AY500769	AY500707	HQ235201	HQ235485
Prunus myrtifolia2004 (L.) Urb.	**Vincent et al.**	HQ188813	HQ243834	HQ244071	HQ244295	HQ235202	HQ235486
Prunus napaulensis422 (Ser.) Steud.	Wen 9277	HQ188814	HQ243835	HQ244072	HQ244296	HQ235203	HQ235487
Prunus nigra040 Ait.	JSh979; var.T: TENN	AY500605	AY500753	AY500772	AY500710	HQ235204	HQ235488
Prunus nigra2016	Wen9910	HQ188815	HQ243836	HQ244073	HQ244297	HQ235205	HQ235489
Prunus oblongum347 Yu & Li	Wen 8441	HQ188816	HQ243837	HQ244074	HQ244298	HQ235206	HQ235490
Prunus oblongum401		HQ188817	HQ243838	HQ244075	HQ244299	HQ235207	HQ235491
Prunus obtusata426 Koehne	Wen 9315	HQ188818	HQ243839	HQ244076	HQ244300	HQ235208	HQ235492
Prunus oleifolia211 Koehne	Wen 53855	HQ188819	HQ243840	HQ244077	HQ244301	HQ235209	HQ235493

Table 9 Continued

Prunus oleifolia409	Nee and Wen 53836	HQ188820	HQ243841	HQ244078	HQ244302	HQ235210	HQ235494
Prunus oligantha2038 Kalkm.	Wen 10743	HQ188821	HQ243842	HQ244079	HQ244303	HQ235211	HQ235495
Prunus orthosepala395 Koehne	DPRU551; RCH202	HQ188822	HQ243843	HQ244080	HQ244304	HQ235212	HQ235496
Prunus ovalis209 Ruiz		HQ188823	HQ243844	HQ244081	HQ244305	HQ235213	HQ235497
Prunus ovalis408	Wen 8625	HQ188824	HQ243845	HQ244082	HQ244306	HQ235214	HQ235498
Prunus padus301 L.	DPRU 1540.1	HQ188825	HQ243846	HQ244083	HQ244307	HQ235215	HQ235499
Prunus padus362	Wen 9028	HQ188826	HQ243847	HQ244084	HQ244308	HQ235216	HQ235500
Prunus pedunculata263 (Pall.) Maxim.	DPRU 23284.4	HQ188827	HQ243848	HQ244085	HQ244309	HQ235217	HQ235501
Prunus pensylvanica049 L.	JSh865; var.T: TENN	AY500632	HQ243849	HQ244086	AY500737	HQ235218	HQ235502
Prunus pensylvanica2041	**JSh 10874; UCHT**	HQ188869	HQ243850	HQ244087	HQ244310	HQ235219	HQ235549
Prunus pensylvanica2043	**JShaw; UCHT**	HQ188870	HQ243851	HQ244088	HQ244311	HQ235220	HQ235550
Prunus pensylvanica2044	**JShaw; UCHT**	HQ188871	HQ243852	HQ244089	HQ244312	HQ235221	HQ235551
Prunus persica053 (L.) Batsch	JSh992; TN: TENN	AY500628	HQ243853	HQ244090	AY500733	HQ235222	HQ235503
Prunus persica274	DPRU 2448	HQ188828	HQ243854	HQ244091	HQ244313	HQ235223	HQ235504
Prunus persica280	DPRU 1469.1	HQ188829	HQ243855	HQ244092	HQ244314	HQ235224	HQ235505
Prunus persica281	DPRU 1474.1	HQ188830	HQ243856	HQ244093	HQ244315	HQ235225	HQ235506
Prunus persica286	DPRU 2275	HQ188831	HQ243857	HQ244094	HQ244316	HQ235226	HQ235507
Prunus persica293	DPRU 2277	HQ188832	HQ243858	HQ244095	HQ244317	HQ235227	HQ235508
Prunus persica299	DPRU 2019	HQ188833	HQ243859	HQ244096	HQ244318	HQ235228	HQ235509
Prunus petunnikowii254 (Litv.) Rehd.	DPRU 2227.6	HQ188834	HQ243860	HQ244097	HQ244319	HQ235229	HQ235510
Prunus petunnikowii369	DPRU 2319.6; EB143	HQ188835	HQ243861	HQ244098	HQ244320	HQ235230	HQ235511
Prunus petunnikowii375	DPRU 2227.7; RCH215	HQ188836	HQ243862	HQ244099	HQ244321	HQ235231	HQ235512
Prunus phaeosticta2020 Maxim	Wen 10820	HQ188837	HQ243863	HQ244100	HQ244322	HQ235232	HQ235513
Prunus phaeosticta354	Liu s.n.	HQ188838	HQ243864	HQ244101	HQ244323	HQ235233	HQ235514
Prunus phaeosticta359	Wen 9425-A	HQ188839	HQ243865	HQ244102	HQ244324	HQ235234	HQ235515
Prunus phaeosticta360	Wen 9425-B	HQ188840	HQ243866	HQ244103	HQ244325	HQ235235	HQ235516

Table 9 Continued

Prunus phaeosticta423	Wen 9425	HQ188841	HQ243867	HQ244104	HQ244326	HQ235236	HQ235517
Prunus pleiocerasus376 Koehne	DPRU 394.1; RCH216	HQ188842	HQ243868	HQ244105	HQ244327	HQ235237	HQ235518
Prunus polystachya353 (Hook.f.) Kalkm.		HQ188843	HQ243869	HQ244106	HQ244328	HQ235238	HQ235519
Prunus polystachya413		HQ188844	HQ243870	HQ244107	HQ244329	HQ235239	HQ235520
Prunus prostrata275 Labill.	DPRU 1629.7	HQ188845	HQ243871	HQ244108	HQ244330	HQ235240	HQ235521
Prunus prostrata393	DPRU 1629	HQ188846	HQ243872	HQ244109	HQ244331	HQ235241	HQ235522
Prunus pseudocerasus384 Lindl.	DPRU 39	HQ188847	HQ243873	HQ244110	HQ244332	HQ235242	HQ235523
Prunus pullei2013 (Koehne) Kalkm.	Potter 081118-09	HQ188848	HQ243874	HQ244111	HQ244333	HQ235243	HQ235524
Prunus pumila L. var. *depressa253* (Pursh.) Bean	DPRU 1939	HQ188849	HQ243876	HQ244113	HQ244334	HQ235245	HQ235525
Prunus pumila059 L.	Horn 2001-02; TN: TENN	AY500623	HQ243875	HQ244112	AY500728	HQ235244	HQ235526
Prunus reflexa215 Walp.		HQ188850	HQ243877	HQ244114	HQ244335	HQ235246	HQ235527
Prunus reflexa216		HQ188851	HQ243878	HQ244115	HQ244336	HQ235247	HQ235528
Prunus reflexa399	Nee and Wen 53868	HQ188852	HQ243879	HQ244116	HQ244337	HQ235248	HQ235529
Prunus reflexa400	Nee and Wen 53820	HQ188853	HQ243880	HQ244117	HQ244338	HQ235249	HQ235530
Prunus rivularis022 Scheele	Endquist 3372; TX: BRIT	AY500597	AY500745	AY500764	AY500702	HQ235250	HQ235531
Prunus rivularis2017	Wen9722	HQ188854	HQ243881	HQ244118	HQ244339	HQ235251	HQ235532
Prunus salicina075 Lindl.	DPRU 791: TENN	AY500617	HQ243882	HQ244119	AY500722	HQ235252	HQ235533
Prunus salicina304	DPRU 2450	HQ188855	HQ243883	HQ244120	HQ244340	HQ235253	HQ235534
Prunus salicina317	DPRU 2460.2	HQ188856	HQ243884	HQ244121	HQ244341	HQ235254	HQ235535
Prunus salicina385	DPRU 384.1; EB77	HQ188857	HQ243885	HQ244122	HQ244342	HQ235255	HQ235536
Prunus salicina421	Wen 9291	HQ188858	HQ243886	HQ244123	HQ244343	HQ235256	HQ235537
Prunus scoparia251 (Spach) Schneid.	DPRU 2224-1	HQ188859	HQ243887	HQ244124	HQ244344	HQ235257	HQ235538
Prunus serotina Ehrh. var. *viride208*	Beck and Estes s.n.	HQ188865	HQ243898	HQ244135	HQ244354	HQ235268	HQ235543
Prunus serotina var. alabamensis 2047 (AKA *Prunus alabamensis* C. Mohr)	Kral 88896; USC A.C. Moore Herbarium	HQ188705	HQ243713	HQ243950	HQ244187	**HQ235068**	HQ235350
Prunus serotina044 Ehrh.	JSh1013; TN: TENN	AY500633	HQ243888	HQ244125	AY500738	HQ235258	HQ235539

Table 9 Continued

Prunus serotina2045	**Jshaw; UCHT**	HQ188873	HQ243889	HQ244126	HQ244345	HQ235259	HQ235552
Prunus serotina2046	**Jshaw; UCHT**	HQ188874	HQ243890	HQ244127	HQ244346	HQ235260	HQ235553
Prunus serotina2049	**Jshaw; UCHT**	HQ188875	HQ243891	HQ244128	HQ244347	HQ235261	HQ235554
Prunus serotina2050	**Jshaw; UCHT**	HQ188876	HQ243892	HQ244129	HQ244348	HQ235262	HQ235555
Prunus serotina2052	**Jshaw; UCHT**	HQ188860	HQ243893	HQ244130	HQ244349	HQ235263	HQ235540
Prunus serotina207	**Jshaw**	HQ188861	HQ243894	HQ244131	HQ244350	HQ235264	HQ235541
Prunus serotina242	**Jshaw**	HQ188862	HQ243895	HQ244132	HQ244351	HQ235265	HQ235542
Prunus serotina345	Wen 7177	HQ188863	HQ243896	HQ244133	HQ244352	HQ235266	HQ235404
Prunus serotina411		HQ188864	HQ243897	HQ244134	HQ244353	HQ235267	HQ235405
Prunus serrulata419 Lindl.	Wen 9858	HQ188866	HQ243899	HQ244136	HQ244355	HQ235269	HQ235544
Prunus sibirica2035 L.	Wen 10665	HQ188867	HQ243900	HQ244137	HQ244356	HQ235270	HQ235545
Prunus simonii074 Carr.	DPRU 545	AY500618	HQ243901	HQ244138	AY500723	HQ235271	HQ235546
Prunus skutchii214 I.M. Johnst	Wen 6828	HQ188868	HQ243902	HQ244139	HQ244357	HQ235272	HQ235547
Prunus sp357	Wen 8755	HQ412798	HQ243903	HQ244140	HQ244358	HQ235273	HQ235556
Prunus sp358	Wen 9025	HQ412799	HQ243904	HQ244141	HQ244359	HQ235274	HQ235557
Prunus sp361	Wen 9004	HQ412800	HQ243905	HQ244142	HQ244360	HQ235275	HQ235558
Prunus spinosa077 L.	DPRU 2289.22: TENN	AY500615	HQ243906	HQ244143	AY500720	HQ235276	HQ235559
Prunus spinosa247	DPRU 848	HQ188877	HQ243907	HQ244144	HQ244361	HQ235277	HQ235560
Prunus spinosa276	DPRU 2399.17	HQ188878	HQ243908	HQ244145	HQ244362	HQ235278	HQ235561
Prunus spinosa285	DPRU 2289.4	HQ188879	HQ243909	HQ244146	HQ244363	HQ235279	HQ235562
Prunus spinosa374	DPRU 473	HQ188880	HQ243910	HQ244147	HQ244364	HQ235280	HQ235563
Prunus spinosissima252 (Bunge) Franch.	DPRU 2226.8	HQ188881	HQ243911	HQ244148	HQ244365	HQ235281	HQ235564
Prunus stipulacea348 Maxim.	Wen 8418	HQ188882	HQ243912	HQ244149	HQ244366	HQ235282	HQ235565
Prunus stipulacea412		HQ188883	HQ243913	HQ244150	HQ244367	HQ235283	HQ235566
Prunus subcordata016 Benth.	J. Syring; CA: TENN	AY500612	AY500760	AY500779	AY500717	HQ235284	HQ235567
Prunus subcordata372	DPRU 2295	HQ188884	HQ243914	HQ244151	HQ244368	HQ235285	HQ235568
Prunus tangutica264 (Batal.) Koehne	DPRU 2327.1	HQ188885	HQ243915	HQ244152	HQ244369	HQ235286	HQ235569
Prunus tenella072 Batsch.	DPRU 2225.6: TENN	AY500629	HQ243916	HQ244153	AY500734	HQ235287	HQ235570
Prunus tenella363	DPRU2225 .11; RCH218	HQ188886	HQ243917	HQ244154	HQ244370	HQ235288	HQ235571

Table 9 Continued

Prunus texana029 Dietr.	JSh 924	AY500611	AY500759	AY500778	AY500716	HQ235289	HQ235572
Prunus tomentosa270 Thunb.	DPRU 2317.6	HQ188887	HQ243918	HQ244155	HQ244371	HQ235291	HQ235573
Prunus tomentosa272	DPRU 2463.1	HQ188888	HQ243919	HQ244156	HQ244372	HQ235292	HQ235574
Prunus tomentosa284	DPRU 0506.7	HQ188889	HQ243920	HQ244157	HQ244373	HQ235293	HQ235575
Prunus tomentosa343	Wen 8059	HQ188890	HQ243921	HQ244158	HQ244374	HQ235294	HQ235576
Prunus tomentosa404	L&W4010 (CS); Cult. CS TS81261	HQ188891	HQ243922	HQ244159	HQ244375	HQ235295	HQ235577
Prunus tomentosa70	DPRU 2316.4: TENN	AY500624	HQ243923	HQ244160	AY500729	HQ235290	HQ235578
Prunus triloba2051 Lindl.	DPRU 2312.1	HQ188892	HQ243924	HQ244161	HQ244376	HQ235296	HQ235579
Prunus triloba368	DPRU 2312.2	HQ188893	HQ243925	HQ244162	HQ244377	HQ235297	HQ235580
Prunus tucumanensis418 Lillo	Nee and Wen 53882	HQ188894	HQ243926	HQ244163	HQ244378	HQ235298	HQ235581
Prunus umbellata Ell. var. *injucunda030* (Small) Sarg.	JSh958-108; GA: TENN	AY500598	AY500746	AY500765	AY500703	HQ235300	HQ235583
Prunus umbellata014 Ell.	JSh774-003; FL: TENN	AY500604	AY500752	AY500771	AY500709	HQ235299	HQ235582
Prunus undulata2022 Buch. -Ham ex D. Don	Wen 10830	HQ188895	HQ243927	HQ244164	HQ244379	HQ235301	HQ235584
Prunus undulata2033	Wen 10656	HQ188896	HQ243928	HQ244165	HQ244380	HQ235302	HQ235585
Prunus undulata259	Wen 8440	HQ188897	HQ243929	HQ244166	HQ244381	HQ235303	HQ235586
Prunus undulata398		HQ188898	HQ243930	HQ244167	HQ244382	HQ235304	HQ235587
Prunus virginiana019 L.	JSh817-040; NH: TENN	AY500634	HQ243931	HQ244168	AY500739	HQ235305	HQ235588
Prunus virginiana2040	JShaw 10854; UCHT	HQ188868	HQ243932	HQ244169	HQ244383	HQ235306	HQ235548
Prunus virginiana2042	Jshaw 10795; UCHT	HQ188899	HQ243933	HQ244170	HQ244384	HQ235307	HQ235589
Prunus virginiana238	JSh871-040	HQ188900	HQ243934	HQ244171	HQ244385	HQ235308	HQ235590
Prunus virginiana378	JSh XXX s.n.	HQ188901	HQ243935	HQ244172	HQ244386	HQ235309	HQ235591
Prunus virginiana429	Wen9906	HQ188902	HQ243936	HQ244173	HQ244387	HQ235310	HQ235592
Prunus wallichii2018 Steud.	Wen 10790	HQ188903	HQ243937	HQ244174	HQ244388	HQ235311	HQ235593
Prunus webbii364 (Spach) Vierh.	DPRU 0197	HQ188904	HQ243938	HQ244175	HQ244389	HQ235312	HQ235594

Table 9 Continued

Prunus wilsonii428 (Diels ex Schneid.) Koehne	Wen 9344	HQ188905	HQ243939	HQ244176	HQ244390	HQ235313	HQ235595
Prunus zippeliana2025 Miq.	Wen 10889	HQ188906	HQ243940	HQ244177	HQ244391	HQ235314	HQ235596
Prunus zippeliana2026	Wen 10902	HQ188907	HQ243941	HQ244178	HQ244392	HQ235315	HQ235597
Prunus zippeliana2027	Wen 10914	HQ188908	HQ243942	HQ244179	HQ244393	HQ235316	HQ235598
Prunus zippeliana2031	Wen 10583	HQ188909	HQ243943	HQ244180	HQ244394	HQ235317	HQ235599

APPENDIX D

OUTLINE AND EXPLAINATION OF ELECTRONIC FILES FOR THIS STUDY

Due to the size and scope of this thesis, a number of files and folders were created to organize the data. Below is an outline of all the electronic files created and used for this thesis. The layout of this appendix is arranged in a hierarchical manner, which mirrors how the files are arranged on the computer. In preparing this appendix there are several broad pieces of information that need to be understood in order for this appendix to be used correctly.

The original *Prunus* dataset consisted of 256 samples however, in order to measure species identification success, only those taxa with two or more samples could be used. This resulted in species without a conspecific pair being removed from the dataset. This left 203 samples of *Prunus*; *Physocarpus opulifolius* was included as an outgroup taxon for Bayesian analyses. It is also present in many of the Uncorrected pairwise distance (UpD) Excel files since the same file created for Bayesian analyses could be used to generate genetic distances in PAUP. This reduced the amount of time spent on creating matrices for Bayesian and UpD analyses. Datasets that have been reduced to 203 or 204 taxa contain 'DeleTaxa' within their file name. There are no analyses files for matrices containing 256 samples. Originally, both *trnLLF* and *trnSGG* were considered single gene regions, respectively, but for publication purposes both were broken apart to accurately measure species identification, i.e. *trnL* intron, *trnL-F* IGS, *trnG* Intron, and *trnG-S* IGS. This separation was only done for the deleted taxa files.

In many cases, there are redundant matrix files owing to the need to create an initial matrix and then create a new file with Bayes block added to the end using Mesquite. These files are the same files with the exception of a Bayes block added to the end and contain 'MrB' within their file name. Importantly, all MrBayes output files contain the original matrix uploaded to Parallel MrBayes Online. This ensures that one can always find the correct file to run future analyses or double-check results. All files have been named using a number and lettering shorthand (ex *matK* = A2) and the gene regions being tested in a particular matrix, like A2, can be found in the file 'DNABC_Analyses Combinations.xls'.

All files are contained in the master file labeled 'Ian Cohen Final DNABC Files'. This folder is located on the hardrive of the Macintosh computer located by the window in the Plant Lab (Holt 115). Within this master file, there are nine subfolders and two word documents. Below is a detailed explanation of the files that are contained within each of these subfolders to help a user navigate through each folder and find the information that they may need.

0. Appendix III—This Microsoft Word file contains a list and explanation of all the files that were created and used for this thesis. It is the electronic version of this document.

1. All Barcode Papers—This folder contains 112 peer-reviewed papers on DNA barcoding, chloroplast genome evolution, and *Prunus* L. (Rosaceae). Not all papers were referenced in this thesis.

2. CBOLInfo&Data—This folder contains four files that correspond to PCR reaction conditions, Plant Working Group taxa used, and data analysis recommendations from CBOL and the PWG.

3. PowerPoints—This folder contains all Microsoft PowerPoint files created for DNABC talks at academic conferences.

4. SEQS-GENBANK—This folder contains seven MacClade files with all the samples that were submitted to GenBank for this thesis project. Files labeled CUT had the front end trimmed before submission to GenBank so that all sequences started at the same nucleotide position.

5. Sequencher Files—This folder contains two subfolders that are organized by the taxonomic scope of the thesis.

- Prunus Sequencher Files—Contains one Excel file, 'A. Prunus-SequencherChecklist.xls,' that corresponds to the samples sent to Macrogen for sequencing. Samples sent to Macrogen were labeled with numbers and in the Excel file, the species name and corresponding number are present to match to the Macrogen files (ex. 133 refers to *Prunus bucharica (Korsh.) Hand.-Mazz.*). There are five subfolders that correspond to the sequencher files, and unaligned MacClade files for each gene region tested. The subfolder 'RAW Prunus Sequences' contains seven subfolders that contain all the original sequences obtained from the Molecular Biological Research Facility, Knoxville, TN or Macrogen Inc., Seoul, Korea. Files are separated by gene region, sequences from samples that did not belong to the genus *Prunus*, and non sequencher files from Macrogen (ex. PDF tracer files or phrd files).
- T&H_DNABC Files—Contains all sequences, sequencher files, and unaligned MacClade files for *matK* and *rbcL*, which were added to the original *Tortoise and the Hare* dataset to compare the utility of these two regions to that of 34 noncoding cpDNA regions.

6. Single_Locus_Aligned_Matrices—Contains the single locus matrices for each gene region tested using samples from the genus *Prunus*. Matrices have been aligned and are either coded using Fast Gap (FG) or not coded at all.

- 204 Samples
    - Singlesw/ofg—This subfolder contains nine files representing the seven single locus gene regions plus *trnLLF* and *trnSGG* concatenated; Gaps are uncoded, hence the file name includes "w/ofg" = without Fast Gap coding.
    - Singlesw/fg— This subfolder contains nine files representing the seven single locus gene regions plus trnLLF and trnGGS concatenated; Gaps have been coded using Fast Gap Program, hence the file name includes "w/fg" = with Fast Gap coding. Note: *rbcL* does not contain any gaps but to make it easier for future users to find the correct data file it was placed in this folder.
- 256 Samples
    - Singlesw/ofg—This subfolder contains five files representing each of the gene regions tested. In this folder, *trnLLF* and *trnGGS* were not separated into the *trnL* intron, *trnL-F* IGS, *trnG* Intron, or *trnG-S* IGS since there were no analyses done for the thesis or publication using the 256 dataset; Gaps are uncoded.
    - Singlesw/fg—This subfolder contains four files representing four of the five single locus gene regions. *RbcL* was excluded because it does not contain any gaps to code. This file can be found in the above subfolder. In this folder, *trnLLF* and *trnGGS* were not separated into *trnL* intron, *trnL-F*

IGS, *trnG* Intron, or *trnG-S* IGS since there were no analyses done for the thesis or publication using the 256 dataset ; Gaps are uncoded (FG = fast gap coding.; Gaps are coded using Fast Gap Program. Note: *rbcL* does not contain any gaps but to make it easier for future users to find the correct file that I used for my thesis work it was placed in this folder.

7. Prunus Matrix combinations based on 204 samples—This folder contains one Excel file that shows which gene regions are specified in each matrix (ex. A3 corresponds to *rbcL*). There are also six numbered subfolders. This represents all the possible combinations of gene regions. Listed below is the number of files in each folder. There are 127 matrices, which represents all the possible gene region combinations. These files are a separated by the number of regions concatenated (e.g., The subfolder labeled 2 contains matrices with two gene regions combined, like *rbcL+matK*).

- 1: Seven files are contained in this folder.
- 2: 21 files are contained in this folder.
- 3: 35 files are contained in this folder.
- 4: 35 files are contained in this folder.
- 5: 21 files are contained in this folder.
- 6&7: Eight files are contained in this folder.

8. Prunus Analyses_Files—These are the data files used to determine species identification success for each gene region alone and in combination. The use of 'DeleTaxa' refers to files with 203 samples or 204 samples depending on if *Physocarpus opulifolius* was included as an outgroup taxon for Bayesian analyses. In some distance matrices, it was not included.

A. DNABC_Analyses Combinations.xls—This file is a list of all 127 possible combinations. Files in the subfolders below are labeled based on position in chart (e.g., *matK* = A2).
B. Coded, Fused, and Bayes Block Matrices—This subfolder contains six folders. Each folder contains the coded and concatenated gene regions for all of the possible gene region combinations (ex. 6 equals 6 gene regions Refer to DNABC_Analyses Combinations.xls file for which gene regions are contained in each file.
C. Multiphyl_DeleTaxa_MrBayes—This subfolder contains the files that were created using the online program Multiphyl to determine which evolutionary model to use for MrBayes. In all, there are nine files representing the seven single locus gene regions plus *trnLLF* and *trnSGG* concatenated, respectively).
D. MrBayes Output Files—This subfolder contains six folders. Each folder contains the coded and concatenated gene regions for all of the possible gene region combinations (ex. 6 equals 6 gene regions combined) and Parallel MrBayes output for Bayesian analysis portion of study. Refer to DNABC_Analyses Combinations.xls file for which gene regions are contained in each file.
E. UpD_Excel_Files—This subfolder contains six folders. Each folder contains the coded and concatenated gene regions for all of the possible gene region combinations (ex. 6 equals 6 gene regions combined). Each folder contains Excel files with genetic distances that were calculated using PAUP for the genetic

distance analysis portion of study. Refer to DNABC_Analyses Combinations.xls file for which gene regions are contained in each file.

9. DNABC_Tables&Figures—This folder contains five subfolders corresponding to tables created for the thesis and publication.

- Bayes Trees—This subfolder contains six Bayesian Tree graphics produced in FigTree v1.3.1 in PDF form.
- MiscTables—This subfolder contains two Excel files, Prunus spp. ID_Checks.xls and PrunusTable1MAY-2010.xls. The first file contains a list of *Prunus* samples with questionable identifications based on DNA barcoding results. The second file contains key information on all of the *Prunus* samples used in this study (ex. Collector, collection number, UTC ID).
- Shaw Files—This subfolder contains the original *Tortoise and the Hare* tables. These tables were not altered in any way but were saved as different versions and subsequently expanded to add *matK* and *rbcL* data. Only the file 'Shaw et al TH3 Appendix S2.xls' was used in my study.
- Tables_For_DNABC_Pub—This subfolder contains three tables to be used for the publishable unit.
- Thesis Tables—This subfolder contains 10 Excel files that were used in this thesis. All files are up to date and correct.

10. IC_Thesis_Rewrite_Figs1-18_JSh.doc—This is the completed thesis file

11. Joey'sPrunusMacCladeFiles—This subfolder contains eight MacClade files from Joey's PhD work. Files were used to determine which taxa sequence data had already been obtained for.

APPENDIX E

CURRICULUM VITA

Ian Cohen
1325 Clearpoint Dr.
Hixson, TN 37343
423-645-4761
Ian-Cohen@mocs,utc.edu

EDUCATION

University of Tennessee at Chattanooga, Chattanooga, TN
M.S. Environmental Science **2008-2010**
Plant DNA barcoding: Testing the utility of the Consortium for the barcoding of Life's Two 'Agreed Upon' Loci

University of Tennessee at Chattanooga, Chattanooga, TN
B.S. Biology **2008**
Cum Laude

AWARDS & HONORS

- Sigma Xi Graduate Student Research Award **2010**
- Association of Southeastern Biologists Graduate Student Support Award **2010**
- University of Tennessee at Chattanooga Graduate Student Association Travel Award **2010**
- UTC Provost Student Research Grant **2009**
- Association of Southeastern Biologists Graduate Student Support Award **2009**
- University of Tennessee at Chattanooga Graduate Student Association Travel Award **2009**
- National Science Foundation- REU Internship **2007**
- UTC Cooperative Education & International Programs Travel Award **2007**
- Geospatial Information & Technology Association Grant **2006**

CLASSROOM EXPERIENCE

Adjunct Faculty **Fall 2009**
Instructor for Introductory Biology I lab in the Dept. of Biological and Environmental Sciences at the University of Tennessee at Chattanooga. Responsible for developing lesson plans for the various lab activities each week, as well as, quizzes and exams. **Spring 2010**

Teaching Assistant **Fall 2005**
Teaching assistant for General Chemistry I and II labs at the University of Tennessee at Chattanooga. Assisted students with experiments and ensured they maintained and followed safe working conditions in lab; also responsible for cleaning up and restocking materials used by students. Worked under, lab instructors Tracy Bingham and C.K. Reynolds. **Fall 2006**

RESEARCH EXPERIENCE

University of Tennessee at Chattanooga, Chattanooga, TN
Thesis **2008-2010**
Plant DNA barcoding: Testing the utility of the Consortium for the barcoding of Life's Two 'Agreed Upon' Loci. Worked with Dr. Joey Shaw on determining the efficacy of DNA barcoding. Generated data for three noncoding chloroplast regions (*trnS-trnG* IGS, *trnG* intron, *trnL-trnF* IGS, *trnL* Intron, and *psbA-trnH* IGS), and portions of two coding chloroplast regions (*matK* and *rbcL*)

University of Tennessee at Chattanooga, Chattanooga, TN

Research Assistant **2008-2010**

Worked for Dr. Stylianos Chatzimanolis, as an NSF funded research assistant, on a multilocus phylogenetic study of the rove beetle tribe Staphylinini (Coleoptera: Staphylinidae). Responsible for generating molecular data and maintaining and organizing all data collected. Project website: http://www.staphylinini.org

University of Tennessee at Chattanooga, Chattanooga, TN

Graduate Independent Study **Fall 2009**

Independent study with Dr. Thomas Wilson that focused on using Arc GIS. Completed online modules through the ESRI tutorial website. Modules completed include: Learning Arc GIS, Creating and Maintaining Metadata using Arc GIS, Creating and Editing Geodatabases for Arc GIS, Arc GIS spatial analyst and Learning Arc GIS analyst.

University of Tennessee at Chattanooga, Chattanooga, TN

Independent Study **Spring 2008**

Independent study with Dr. Joey Shaw focused on understanding how to and learning about molecular genetics/systematic lab techniques. Met for three hours a week and learned how to maintain accurate records, extract plant DNA, set-up and perform PCRs, gel electrophoresis, ExoSAP-IT protocol, ABI Big Dye sequencing reaction, and Sephadex DNA purification protocol.

University of Puerto Rico, Rio Piedras. Puerto Rico

National Science Foundation-REU Intern **2007**

Field study with Dr. James D. Ackerman, focused on *Oeceoclades maculata* (Lindl.) Lindl. (Orchidaceae), an invasive species found throughout the Neotropics. Program involved developing a research project, oral presentations, gathering field data, reviewing literature, and writing a manuscript. Manuscript published in 2009. *Annals of Botany* 104: 557-563.

Tennessee Dept. of Environment and Conservation, Chattanooga, TN

Field Surveyor/ Botanist **2007**

Served as the botanist for a field crew. Primary responsibilities were to find, positively identify, and count the occurrences of *Scuttelaria montana* Chapm. (Lamiaceae) in several state natural areas in northern Hamilton County, Tennessee. Goal was to help stop rock/mineral harvesting in and around protected areas.

University of Tennessee at Chattanooga, Chattanooga, TN

Independent Study **2006**

"A Preliminary Floristic Survey of the Tennessee River Gorge Grant Tract, Marion County, Tennessee." Study was in conjunction with the Tennessee River Gorge Trust, a nonprofit conservation group. Project included collecting, identifying, and preparing specimens to be placed in the University of Tennessee at Chattanooga Herbarium (UCHT). Learned how to prepare a manuscript and presented a poster at the 2007 Annual Meeting of The Association of Southeastern Biologists, Columbia, SC 2007 and The Annual Meeting of The Tennessee Academy of Sciences, Gallatin, TN 2007.

PAPERS

- Chatzimanolis, S., **I.M. Cohen**, A. Schomann, and A. Solodovnikov. 2010. *Towards a robust phylogeny of the mega-diverse rove beetle tribe Staphylinini: Molecular data and their evaluation* (Insecta, Coleoptera, Staphylinidae). *Zoologica Scripta* 39: 436-449.
- **Cohen, I.M.** and J.D. Ackerman. 2009. *Oeceoclades maculata*, a tropical orchid invades a Caribbean rainforest. *Annals of Botany* 104: 557-563.

PRESENTATIONS

- **Cohen, I.M** and J. Shaw. 2010. *Barcoding: Testing the Utility of One Coding and Three Noncoding Chloroplast DNA Regions using Prunus L. (Rosaceae) as a model*. Paper presentation, Annual Meeting of the Southeastern Biologists, Asheville, NC.
- Shaw, J., J. Wen, **I.M. Cohen**, R. Haberle, C. Siew-Wai, and D. Potter. Chloroplast DNA phylogeny of *Prunus* L. (Rosaceae) using *trnS-trnG-trnG*, *psbA-trnH*, *trnL-trnL-trnF*, and *matK* cpDNA Sequences. Paper presentation, Annual Meeting of the Southeastern Biologists, Asheville, NC.
- **Cohen, I.M**. 2009. Testing the Utility of Three Noncoding Chloroplast DNA Regions for DNA Barcoding using Prunus (Rosaceae) as a Model: Trials and Tribulations of an Inexact Science. Invited speaker, undergraduate seminar class, University of Tennessee at Chattanooga, Chattanooga, TN,
- Chatzimanolis, S., **I. Cohen**, A.M. Schoman, and A. Solodovnikov. 2009. A preliminary multi locus phylogeny of Staphylinini (Coleoptera: Staphylinidae). Paper presentation, Annual Meeting of the Entomological Society of America, Indianapolis, IN.
- **Cohen, I.M.** and J. Shaw. 2009. Testing the Utility of Three Noncoding Chloroplast DNA Regions for DNA Barcoding using Prunus (Rosaceae) as a Model. Paper presentation, Annual Meeting of Tennessee Academy of Sciences, Knoxville, TN.
- **Cohen, I.M.** and J. Shaw. 2009. Testing the Utility of Three Noncoding Chloroplast DNA Regions for DNA Barcoding using Prunus (Rosaceae) as a Model. Paper presentation, Annual Meeting of The Association of Southeastern Biologists, Birmingham, AL.
- **Cohen, I**. and J. Shaw. 2007. A preliminary floristic survey of the Tennessee River Gorge Grant Tract, Marion County, Tennessee. Poster presentation, Annual Meeting of The Tennessee Academy of Sciences, Gallatin, TN.
- **Cohen, I**. and J. Shaw. 2007. A preliminary floristic survey of the Tennessee River Gorge Grant Tract, Marion County, Tennessee" Poster presentation, Annual Meeting of The Association of Southeastern Biologists, Columbia, SC.

PROFESSIONAL MEMBERSHIPS

- Association of Southeastern Biologists
- Botanical Society of America

LANGUAGES

- English – Native language
- Spanish – Speak & read with medium proficiency; basic writing skills.

REFERENCES

- Dr. Joey Shaw

University of Tennessee at Chattanooga
Department of Biological and Environmental Sciences
215 Holt Hall
Dept 2653
615 McCallie Avenue
Chattanooga, TN 37403
Office Phone: 423-425-4265
Joey-Shaw@utc.edu

- Dr. Stylianos Chatzimanolis

University of Tennessee at Chattanooga
Department of Biological and Environmental Sciences
215 Holt Hall
Dept 2653
615 McCallie Avenue
Chattanooga, TN 37403
Office Phone: 423-425-4341
Stylianos–Chatzimanolis@utc.edu

- Dr. James D. Ackerman

University of Puerto Rico, Rio Piedras
Department of Biology
NCN Fase I, Room 136
Barbosa Avenue
San Juan, Puerto Rico 00931-3360
Office Phone: (787) 764-0000 ext. 2023, 4851
(787) 764-0000 ext. 2901
Ackerman.upr@gmail.com

- Linda Collins

University of Tennessee at Chattanooga
Department of Biological and Environmental Sciences
215 Holt Hall
Dept 2653
615 McCallie Avenue
Chattanooga, TN 37403
Office Phone: 423-425-4797
Linda-Collins@utc.edu